

## DOCUMENT RESUME

ED 280 463

IR 051 822

**TITLE** IFLA General Conference, 1986. Pre-Conference Seminar on Automated Systems for Access to Multilingual and Multiscript Library materials: Problems and Solutions. Papers.

**INSTITUTION** International Federation of Library Associations and Institutions, The Hague (Netherlands).

**PUB DATE** Aug 86

**NOTE** 281p.; Papers presented at the International Federation of Library Associations (IFLA) General Conference, Pre-conference Seminar (52nd, Tokyo, Japan, August 21-22, 1986).

**PUB TYPE** Speeches/Conference Papers (150)

**EDRS PRICE** MF01/PC12 Plus Postage.

**DESCRIPTORS** \*Bibliographic Utilities; \*Databases; Developed Nations; Developing Nations; Foreign Countries; \*Information Systems; International Organizations; Library Associations; Library Automation; \*Library Materials; \*Multilingual Materials; Online Catalogs; \*Online Systems

**IDENTIFIERS** Bibliographic Data Bases; \*Nonroman Scripts; Transliteration

**ABSTRACT**

A seminar which considered problems and solutions regarding automated systems for access to multilingual and multiscript library materials was held as a pre-session before the IFLA general conference in 1986. Papers presented include: (1) "Romanized and Transliterated Databases of Asian Language Materials--History, Problems, and Prospects" (Suzine Har-Nicolescu, United States); (2) "Non-Roman and Multi-Script Bibliographic Databases--Basic Issues in Design and Implementation" (Alan M. Tucker, United States); (3) "Computer Processing of Southeast Asian Scripts, with Particular Reference to Jawi and Thai" (Edward Lim Huck Tee, Malaysia); (4) "Use of Arabic Script in Computerized Information Systems" (Mohammed M. Aman, United States); (5) "Development of Greek Automated Cataloguing System--Victoria, Australia" (Eve Stocker, Australia); (6) "Character Set and Its Control in an Online Cataloguing System--Application at the Center for Bibliographic Information" (Akira Miyazawa, Japan); (7) "Utlas Developments with Japanese and Chinese Scripts" (Jack Cain, Canada); (8) "The CJK (Chinese, Japanese, and Korean) Enhancements to the RLIN (Research Libraries Information Network) System--A Review of Basic Issues" (John W. Haeger, United States); (9) "Bibliographic Control and Library Automation in Japan" (Shojiro Maruyama, Japan); (10) "A Universal Coding System for the Multi-Lingual Environment" (S. S. Tseng, C. C. Hsieh, Jack K. T. Huang, C. T. Chang, and C. C. Yang, Taiwan); (11) "What Next? Issues Arising from Conference Deliberations" (Cynthia J. Durance, Canada); (12) "The Review of the International Standard Bibliographic Descriptions (ISBDs) and Non-Roman Script Publications" (Barbara Jover, United Kingdom); (13) "MINISIS--A Multilingual Information Management System" (Richard Lee, Canada); and (14) "Computers in a Multi-Script Library Environment" (Sylva Simsova and John Clews, United Kingdom). (KM)

☒ This document has been reproduced as  
received from the person or organization  
originating it.

☐ Minor changes have been made to improve  
reproduction quality.

• Points of view or opinions stated in this docu-  
ment do not necessarily represent official  
OERI position or policy.

IFLA

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS FOR  
ACCESS TO MULTILINGUAL AND MULTISCRIP LIBRARY  
MATERIALS: PROBLEMS AND SOLUTIONS.

Tokyo, Japan

August 21-22, 1986

BEST COPY AVAILABLE

"PERMISSION TO REPRODUCE THIS  
MATERIAL HAS BEEN GRANTED BY

A.L. Van Wesemael

TO THE EDUCATIONAL RESOURCES  
INFORMATION CENTER (ERIC)."

ED280463

IR051822





1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

61-MCULTP-1-E  
INF

ROMANIZED AND TRANSLITERATED DATABASES OF ASIAN-LANGUAGE MATERIALS:  
HISTORY, PROBLEMS, AND PROSPECTS

by

Suzine Har-Nicolescu

Chief Librarian and Professor, Library Division

Medgar Evers College/City University of New York;

Trustee, Brooklyn Public Library

Brooklyn, New York, U. S. A.

Keywords:

Universal Bibliographic Control (UBC)  
Asian languages  
Transliteration  
Romanization  
Library Automation

For internal use only:

To be brought to meeting no:

----- 21 August '86 -----

Remarks:

to be presented orally ☐

for distribution only ☐

西欧における自動化された統一的书誌調整の現状は、統一的な目録標準規格の統合化によって、アジア諸国におけるよりはるかに進歩したものとなっている。アジアでは、西欧において成功したシステムを適用することによって個々の書誌システムを開発する段階にある。これはアジア諸国すなわち中国、日本ならびに韓国と西欧との間の書誌的な垂直的交流の結果である。しかしながら地理的な範囲と「書きもの」において共通する文字の使用によって歴史的に、中国、日本ならびに韓国(CJK)を偉大な文化的な縁(えにし)によって結びつけているという理由から、アジア諸国間の平等かつ水平的な関係が、現在極めて強く望まれているのである。

これらの国々の図書館は、それぞれ、大きな比率で西欧資料を所蔵しているが、アジア資料に対するこの比率は、将来的には減少することになる。アジア資料の増加と偉大な文化的な紐帯は、アジアの3ヶ国が標準アジア文字セット a standard "Asian" character set を確立すること、地域的ネットワークを確立することを必然のものとしている。

国際標準書誌記述(ISBD)に基づいたUNIMARCフォーマットならびに英米目録規則第2版(AACR 2)によって目録業務を自動化する活動は、すでにアジア諸国において受入れられているところである。しかしながら、このシステムは、OCLCやRLINのような自動化した目録作成システムと同様に、主に西欧の顧客むけに開発されたものである。最近に至り、MARCフォーマットならびに標準ローマ字化表の開発を通じてアジア諸語、特にCJKのための自動化目録作成システムが出現しはじめた。しかしながら、これらのシステムの開発において、アジア各国はそれ程本格的に相談を受けたという訳ではない。MARCフォーマットの目録カードでは、その過度の句読記号は、アジアにおける要件としては不適当であり、また魅力的ではない。また標題紙志向の目録作業は、アジア人にとってもわずらわしいものである。UNIMARCフォーマットは、アジア各国の目録作業の短所である主題アクセスのためには十分の手当がされていない。

現在稼働中のアジア資料用の自動化目録作成システム—RLINの"Sinoterm"システム、OCLCの"Asiagraphics"—においてもデータ入力の問題がある。これらの問題は、コンピュータのキーボード上での文字の複雑性、複雑な構造(シンタックス)と形(モルフォロジー)に起因するアジア諸文字のローマ字化の困難性などからひき起こされている。中国語、日本語そして韓国語の文字をコンピュータに読み込むためのプログラミングの際に、アジアの図書館員自体も十分に相談にあずかってはいない。

アジア3ヶ国の言語間での(すなわち漢字の適用)における構文ならびに形態上の類似と差異とは、絶好の機会と大きな挑戦なのである。例えば、3ヶ国を合せて、各国のローマ字化表を組み合わせれば、日本人著者の名前は少なくとも12通りにもローマ字化できる。これらの問題は、書誌的な統一要件を確立する上で障壁となっている。要するに、アジア人は、ローマ字化を西欧の利用者のための手段、経済的には非現実、アジア人の利用者には概念的には不適当と考えているのである。

結論と勧告として：アジア諸語と西欧諸語の重大な差異を理解すべきである；アジア各国はまず各自の利用者に提供するための情報システムを開発することのメリットを認識すべきである；ローマ字化を西欧人の利用者にサービスするオプションと考え維持すべきである；アジア各国に対して各自の利用者向けのアジア資料に適した書誌調整を開発することができるよう、十分に具備された指針と支援を提供すべきこと；東洋ならびに西洋の双方から等しい数の図書館員と機械化担当者を招集して、ユニバーサルな書誌的規格の開発を促進すべきである；そして東西の意見交換の機会を増すべきこと；最後にIFLAは開発途上国からの参加に対して財政援助を提供し、またこの提案によって情報システムの世界の中での東洋各国の位置を認識すべきこと。

## ABSTRACT

S. Har-Nicolescu (USA). Romanized and transliterated databases of Asian-language materials: history, problems, and prospects.

This article addresses the status of universal bibliographic control with special attention to China, Japan, and Korea (CJK).

International Standard Bibliographic Description (ISBD) has been developed with insignificant consultation with Asian representatives - especially those librarians of China, Japan, and Korea. We must note also that each of those countries is in the process of developing its own bibliographic system through the ISBD/MARC format. The problems resulting from these isolated independent bibliographic developments are further complicated by the problems of transliterating Asian languages through romanization, of transliterating between Asian languages, and of adapting transliteration schemes to computer protocols. This article reviews these problems of transliteration and offers an historical perspective on the origins of Asian writing systems, especially those of CJK.

The conclusion and recommendations in this paper pertain to the need for a cooperative regional Asian bibliographic network, the responsibilities of librarians and automators in Asian and Western countries, and IFLA's role in the equitable and scholarly development of standards for international bibliographic exchange.

ROMANIZED AND TRANSLITERATED DATABASES OF ASIAN-LANGUAGE MATERIALS:  
HISTORY, PROBLEMS, AND PROSPECTS

More than ever before the technological revolution has brought the nations of the world closer together intellectually. With the proliferation of computers, global exchange and dissemination of information through cooperative resource-sharing networking have become increasingly desirable, particularly between English and non-English speaking communities. To achieve such a goal, however, all type of library materials, regardless of language or origin, must be controlled bibliographically through the implementation of uniform standards of cataloging and encoding.

In the West, such bibliographic uniformity is currently much more highly developed than in Asian countries. Within this context, I would like to offer several observations on present problems and future possibilities in international bibliographic control.

I. POLITICAL/ECONOMIC/CULTURAL PROBLEMS AND OPPORTUNITIES

Until recently, political and economic realities in China, Korea, and Japan have pushed each country towards separate interests and towards the development of separate bibliographic systems for patrons. If this trend were to continue, each country will develop unilaterally its own automated bibliographic system by adapting in its own fashion those systems seemingly successful in the West. This independent interaction among each of the three Asian countries and the Western world without an equally interdependent relationship between Asian countries is, in my opinion, only a transitional phenomenon; and it does not offer a long-

term solution for the benefit of Asian countries.<sup>1</sup>

While each of the three countries is proud of her own unique heritage, geographical proximity and the use of common characters in their writing systems have, historically, bound the three countries together with major cultural ties. Since each country is striving to meet its national and international information priorities, together the three countries at this date have not been successful in establishing a mutually agreed-upon set of regional standards for transliterating their languages and for the introduction of other information products and services for Asian countries as a whole.

Western materials (English and other roman-alphabet materials) comprise a large portion of the foreign-language collections in the libraries of China, Korea, and Japan. It is most unlikely, however, that the Asian materials in each of the libraries of these countries would be fewer than the volume of Western holdings. On the contrary, owing to steady economic growth and advancements in modern technology, the publishing industries in all three countries have been growing. This will keep the collections more heavily Asian than Western in the libraries. This trend certainly shows no sign of slackening.<sup>2</sup>

To create an environment hospitable to such a development, it is imperative that the three Asian countries come together to establish a standard "Asian" character-set that would promote a mutual exchange of information.<sup>3</sup> Such a move would serve as a first step to the realization of regional networking. This regional Asian network would then be able to establish a strong and united information front with which



to interact with Western bibliographic networks, and would also create a far more efficient and effective bibliographic enterprise than any currently in operation in the Asian world.

## II. INTERNATIONAL BIBLIOGRAPHIC STANDARDS & SYSTEMS: BACKGROUND

Following the creation and development of (LC)MARC (Library of Congress Machine-Readable Cataloging) in the mid 1960's and through the international exchange of MARC records (UNIMARC), the Universal Bibliographic Control (UBC) Program was made possible.<sup>4</sup> Thereafter, many countries gradually adopted to the ISBD (International Standard Bibliographic Description) based UNIMARC format<sup>5</sup> along with AACRII (Anglo-American Cataloging Rules II) to automate cataloging activities. By 1981, over ten countries, including several Asian nations, were following the new format and new cataloging rules.<sup>6</sup>

In the mid 1970's, several automated cataloging systems were developed, and romanization schemes were standardized to control non-roman language materials. By 1981 there were the Bibliographic Automation of Large Library Operations Using Time-Sharing System (BALLOTS, now called RLIN), the Ohio College Library Center (OCLC, now called the On-Line Computer Library Center), the Washington Library Network (WLN, now called the Western Library Network), and the University of Toronto Library Automated Systems (UTLAS). As the data storage capacity of these systems expanded at an accelerated rate, nation-wide information resource-sharing networking became a reality. But these systems were developed mainly for Western languages.

Through the development of the MARC format and a standard romani-

zation scheme, automated cataloging systems for Asian-language materials - especially, Chinese, Japanese, and Korean (CJK) - appeared. In 1983, RLG's RLIN, in cooperation with LC, installed on a trial basis a CJK system in ten major academic and public libraries to catalog Asian materials.

At almost the same time, OCLC developed a similar system, Asia-graphics. Just a few months ago, OCLC announced the completion of a CJK system with expanded capacity to catalog Chinese, Japanese, and Korean materials.

UTLAS (in Toronto) has also been very active in developing its automated cataloging system and has already customized its services primarily in Japan.<sup>7</sup> RLIN and OCLC are applying, and UTLAS will apply, in similar ways a standard romanization scheme to catalog Asian materials.

There is no doubt in my mind that these systems will continue to expand their services in Asian countries in the years ahead. But the Asian countries themselves (and certainly their librarians) were not adequately consulted or involved in the development of these systems.<sup>8</sup>

### III. BIBLIOGRAPHIC CONTROL OF ASIAN MATERIALS: PROBLEMS

#### A. ISBD/MARC

The purpose of ISBD is to promote the "international communication of bibliographical information." The introduction to the 1971 edition of ISBD states:

By specifying the elements which should comprise a bibliographic description and by prescribing the order in which they should be presented and the punctuation by which they should be demarcated, it aims at three objectives: 1) to make records from

different sources interchangeable, 2) to facilitate their interpretation across language barriers, and 3) to facilitate the conversion of such records to machine-readable form.<sup>9</sup>

The method adopted for achieving these objectives is the development of standard punctuation symbols used as "tags," or "delimiters," in order to facilitate the interpretation of bibliographic records across language barriers.

It is quite questionable in my mind, however, whether ISBD will ever achieve any of its objectives, for the reasons Gerald Swanson so aptly explains.<sup>10</sup> If Kevin Hegarty finds MARC unfriendly to the Western patron (with the exception of trained catalogers),<sup>11</sup> I would say that MARC is most assuredly hostile to Asian patrons. While MARC punctuation may be somewhat helpful for Western patrons lacking a knowledge of Asian scripts, this punctuation enormously confuses patrons who read and understand Asian scripts.

The reasons are historic. Chinese characters still form the integral foundation of both the Korean and Japanese writing systems. These characters are pictographic and/or ideographic in nature. Each character denotes a clear meaning, and the number of meanings multiply as symbols increase in number and combination. For example, without knowing how to pronounce the Japanese language, young Korean students can fairly well grasp the general range of meanings of a Japanese sentence written in Kanji or vice versa.

Catalog cards in MARC format seem equally inadequate to Asian needs. Only in the future will on-line catalogs be popular with Asian patrons, so that for now Asian libraries must depend on

catalog cards or printouts. Repetition of the author statement, and the inclusion of multi-lingual titles and sub-titles in the body of the entry confuse Asian patrons. Even more confusing are the slashes, equals-signs, colons, and other punctuation together with diacritical marks built into each romanization scheme.<sup>12</sup> The total effect of all these signs is to produce catalog cards totally unattractive to the aesthetic sensibilities of Asians.

Cataloging rules that emphasize titles are, in my opinion, as cumbersome as is author-oriented cataloging. For CJK librarians the onerous task of transcribing, transliterating, or translating titles requires no fewer, and often more, rules and exceptions than are required for transliterating an author's name. This is particularly true in the case of Asian titles. Often, titles that are too general or deliberately ambiguous require much subtler handling than is possible through romanization. Finally, when looking for a book, the average patron (including the Asian patron) does not think in terms of a title without associating it with an author.<sup>13</sup>

No less important for the UNIMARC format is its insufficient provision for subject access. As Kevin Hegarty points out, a recent study of the use of the on-line catalog shows that most patrons want subject access and that patrons use subject access far more than previously thought.<sup>14</sup> As patrons become more sophisticated in their information demands, without greater subject access they will become even more frustrated.<sup>15</sup>

In this respect, the UNIMARC format is not satisfactory for Asian countries. China, Japan, and Korea service their patrons

with only meager or even no subject access in a true sense, as these countries largely depend on localized DDC-based subjects.<sup>16</sup> The lack of standardized subject headings on the national or regional level is the biggest problem in the cataloging systems of Asian countries.

B. Automated Cataloging Systems for Asian Materials

As a result of intensive research and experimentation, two major automated systems for cataloging Asian materials appeared in 1983: a CJK system (Transtech system), introduced by the Research Library Information Network (RLIN) of RLG;<sup>17</sup> and a CJK system (Asiagraphics), introduced by OCLC.<sup>18</sup> Together, the two systems marked a first step in revolutionizing the automation of non-roman scripts, and in this way helped bridge the gap between Western and Asian bibliographic control and access. However, once again Asian librarians, scholars, and automators were not consulted to a large extent.

At present, there are two available methods of computerization for Asian languages: the "Character-based" (or "Word-decomposition method") used by RLIN, and the "Pronunciation-based" (or "Romanization method") of OCLC.

Though it is too early to evaluate the merits of the two computer systems, some shortcomings are already evident. In RLIN, the method of searching for words in a "Sinoterm" terminal lacks a defined standard that may be referred to by all operators. (A Sinoterm terminal offers the elements necessary for composing Chinese characters.) Specifically, to use the terminal effectively the operator must not only memorize the positions of numerous keys,

he/she must also be thoroughly knowledgeable in the method of "word-decomposing." High production costs and a large keyboard are the remaining shortcomings of the Transtech CJK system.<sup>19</sup>

The Asiagraphics system has simplified, in comparison, the above method by applying a "pronunciation-based" approach. All Chinese characters are input using roman-alphabet keys, with the output showing Chinese characters. In order to internationalize its market, Asiagraphics supports a variety of romanization scheme for Asian languages: the Wade Giles's (a traditional scheme) and the Pin Yin (a new scheme developed by PRC) for Chinese; a modified Hepburn scheme for Japanese; and the McCune-Reischauer scheme for Korean. The "Tsang-chieh" system, the only character-based input method, is used for generating all Chinese characters in both full and simplified forms, whether in Chinese, Japanese Kanji, or Korean Hancha.<sup>20</sup>

Asiagraphics contends that anyone who knows its romanization scheme and the English-language keyboard can operate the system without a knowledge of Chinese characters per se. However, the operator who cannot pronounce a Chinese word will not be able to input data properly. Even with operators who know Chinese, input may vary because pronunciation of the language varies according to dialect. For this reason it is extremely difficult to maintain the integrity of the authority files for databases. At the same time, the operator must always depend on the Chinese word-form or word-meaning unless the problem of distinguishing between homonyms (words of different meanings with the same sound) has been thoroughly resolved.<sup>21</sup>



In other words, regardless of the method of input, both systems must depend heavily on the operator's strong knowledge of Asian languages and English. As long as there is an ample supply of qualified librarians, we are safe. But what would happen if those librarians become scarce or not available at all?

For some time, several information vendors, including OCLC and UTLAS (of Canada), have actively sought a market in the Asian information-world. Equally obvious are the increasing numbers of exchange librarians from the West who visit Asian countries for a short period and then suggest quick solutions to Asian library problems. I personally find healthy such an involvement with Asian nations, for the advanced information technology of the West coupled with a modest degree of guidance will stimulate Asian countries to cooperate with these vendors and librarians so as to establish an information network suitable to each nation and also to all Asian nations.

I would hope, however, that these vendors and librarians keep several issues in mind: 1) how to interrelate, within guidelines set by the individual nations, the Chinese characters used with significantly different pronunciations in the Japanese and Korean writing systems; 2) how to justify philosophically and economically the application of a romanization scheme for Asian languages used by Asian patrons; 3) how to allow Asian libraries to interface their own bibliographic systems without the intermediation of romanization; and 4) how to understand, support, and cooperate with Asian libraries so that the latter can integrate various existing systems (as is

successfully done in Japan)<sup>22</sup> in order to develop a cost-effective system mutually beneficial to Asian and Western patrons.

C. Transliteration of Asian Languages: Romanization

To facilitate universal bibliographic control the Library of Congress, with the cooperation of IFLA, issued standard romanization schemes for the transliteration of non-roman scripts. Since 1976 these romanization schemes have been applied to transliterate non-roman names and texts.<sup>23</sup> This has enabled bibliographic records (both in roman and non-roman script) to be encoded under a single alphabet system. This has also accelerated the development of on-line cataloging systems for Asian materials (Chinese, Japanese, and Korean), extended UNIMARC markets in Asian countries, and facilitated online cataloging systems for CJK materials in the Western world.

Owing to the complex structure (syntactics) and formations (morphology) of Asian languages (complexities compounded by the difficulty of integrating the CJK language systems), computerization of Asian-language materials has been seriously delayed. While more and more librarians in the Western world have participated with enthusiasm in library-automation activities, librarians responsible for Asian-language materials (as I have previously stated) have remained uneasy bystanders to this revolutionary event in librarianship.

Even at this date, many, if not all, catalogers of Asian-language materials continue to use traditional techniques (namely,

handwriting) for the total or partial preparation of catalog cards and other library records in Asian script. The reason for this is not that the needs of computerizing Asian-language materials have been ignored by librarians or automators in the West. The problem lies in the extreme difficulty of programming computers to read Chinese, Japanese, or Korean script. Computers more easily read a roman alphabet and arabic numerals. This difficulty has led to and maintains the wide gap between the automation status of roman-alphabet and non-roman-alphabet materials for the purpose of resource sharing.

Regardless of the apparently obvious differences between the languages of the three Asian countries - China, Japan, and Korea - there are nonetheless a great many syntactical and morphological similarities between the languages of these three countries. The single and most significant factor in these similarities has been the adoption of Chinese characters by Korea and Japan.<sup>24</sup> (See Appendix A.) In adopting Chinese script both Korea and Japan retained and maintained unaltered the original Chinese meanings. This holds true even today and will probably continue to hold true in the future. Pronunciation of the scripts changed, however, slowly and steadily over the centuries: Korea integrated Chinese (Hancha) within its own phonetic alphabet (Hangul), while Japan integrated Chinese script (Kanji) within its own phonetic alphabet (Kana).

Korea resisted any drastic changes in the formation of Chinese characters, and it retained pronunciation closer to that of the original Chinese characters. Japan, on the other hand, freely

changed, added to, deleted from, and distorted the pronunciation of each Chinese character into a multi-sound mode that changes according to time, place, and mood. It is an incredibly difficult task even to the most learned linguist or scholar to ferret out the number of possible different pronunciations of a Chinese character. Every Chinese character in the Japanese reading system denotes an average of two to three sounds; some characters offer as many as eight to ten variations in pronunciation according to the context in which the character appears.<sup>25</sup> (See Appendix B.)

To complicate the situation further, each of the three Asian countries applies many different romanization schemes to transliterate the names of authors or texts: China has three schemes; Korea has three schemes with three variations (due to the various ways of grouping words); and Japan has three schemes. By taking all three countries together and combining all their romanization schemes, the name of a Japanese author can be romanized in as many as twelve different ways. (See Appendix C.)

This is, of course, only the case of transliteration through romanization. How about the case of transliteration into the Japanese language (Japanization) of a name or word from Chinese or Korean? How about the case of transliteration into the Korean language (Koreanization) of a name or word from Chinese or Japanese?<sup>26</sup> (See Appendix D.) Add another six or more possibilities to the previous number of twelve romanized versions.

This leads to an important question: What happens to a book, for example, written by a Japanese author, when one copy is cata-

logged by an American library while an identical copy is cataloged by a Korean library? How can both catalogers arrive at the same bibliographic profile? It is indeed a life-time challenge for anyone!

For these reasons, Asians consider romanization a device to serve Western patrons. For Asians, romanization serves only privileged people - namely, trained librarians and educated patrons. Thus, from the perspective of Asian librarians, romanizing vast Asian collections for the sake of Western patrons is economically unrealistic and philosophically unsound.

#### IV. CONCLUSION AND RECOMMENDATIONS

##### A. General Suggestions

1. We should understand the profound differences between Asian and Western languages, and between Asian and Western peoples. Statements about "the serious language barrier" or "the most thorny problems in information transfer," etc., represent the opinions of Westerners and are not necessarily shared by Asians.
2. We should recognize the merits of information systems developed by Asian countries to serve first their own people. In order to achieve universal bibliographic products and services, we must seek ways to utilize these systems.
3. We should retain romanization schemes as an option for cataloging non-roman script materials for Western patrons. For Asian patrons who do not require romanization schemes,



we should seek an alternative method for standardizing the organization and control of Asian materials.

4. We should address the question of how in equal fashion Western and Asian schemes for the retrieval of and access to cataloged materials should proceed at this time.

5. We should provide guidance to and assistance for Asian countries to allow them to develop a viable and economical system for controlling all Asian-language materials that serve Asian and Western patrons.

6. We should explore the possibility of developing greater subject-access on national, regional, and international levels.

#### B. Application

We should establish and develop a tri-level information system.

As a first step, we should establish the pre-eminent priority of local libraries serving local patrons.

As a second step, we should establish regional bibliographic standards. The Asian countries - especially, China, Korea, and Japan - should together develop a standard "Asian" character-set, and other related products and services in order to establish an Asian bibliographic resource-sharing network.

As a third step, we should work towards an international application by working towards the interfacing of this Asian bibliographic network with that of the West.

Each of the above three steps should be developed without



losing sight of the other two steps so that all three systems can co-exist harmoniously.

C. Genuine International Cooperation

Asian librarians recognize the value of UBC standards and their applications. Western nations should recognize the importance of the efforts of Asian nations to establish information systems best suited to their own patrons and collections along with systems that do not ignore the needs of international resource-sharing.

This requires co-equal, not prejudiced, participation in the development of universal bibliographic standards. Participation by equal numbers of librarians and automators from East and West would be helpful in covering all aspects of information systems.

Often, computer experts tend to emphasize the hardware and software aspects of a problem; linguists emphasize encoding methods; and, with monetary profit in mind, database vendors eagerly seek markets. All these groups tend to neglect the special needs and requirements of library automation, and the technical questions involved in adapting existing automation systems to serve individual countries, regional groupings, and the world-wide information community.

D. IFLA's Role

There is a great need for a coordinated program that will allow meetings or organizations, national and international, to provide more opportunities for an exchange of views and suggestions.<sup>27</sup> IFLA can play, in my opinion, a very important role

in this respect. Seeking financial support to assist the growing number of new participants from developing countries<sup>28</sup> and recognizing the place of Asian nations within the main stream of the world of information systems should be an important priority on IFLA's agenda.

I believe that the welfare of an individual nation should and can coincide with that of regional and global communities. I also believe that this statement applies as well to information systems and services.

The End

## REFERENCES

1. Har-Nicolescu, Suzine, "Cataloging/Coding Systems Employed by Yonsei, Ewha, and Sogang University Libraries in Korea: A Study of the Feasibility and Desirability of Employing Korean Cataloging-Information in Local and International Shared Information Networks." (Fulbright Research Study, Feb., 1985). (To be published.) This report strongly supports the above statement regarding Asian bibliographic trends. Korean and Japanese librarians and automators firmly believe that, though they respect the technological innovations of the Western information-world, their own particular language system should not be supplanted by romanization. An equally strong desire is that Asian countries - especially China, Korea, and Japan - should together develop a standard "Asian" character-set, and other related products and services in order to establish an Asian bibliographic resource-sharing network.
  
2. Har-Nicolescu, "Cataloging/Coding." The report reveals the following: The three university libraries of Korea - Yonsei, Ewha, and Sogang - presently hold Asian and Western materials in the ratio of 60:40. With exception of Seoul National University, these institutions hold the largest library collections in Korea. From 1951 to 1980 the purchasing rate for Asian materials by Yonsei, Ewha, and Sogang had risen by respectively 350%, 525%, and 1400%, while the purchasing rate for Western materials had maintained only moderate growth, in comparison to the Asian materials, by respectively 55%, 33%, and 100%.

Also, as early as 1977 the Korean publishing industry was ranked 15th in the world. (A Handbook of Korea, 3rd ed. Seoul: Korean Overseas Information Service, Ministry of Culture and Information, 1979, p. 715.)

Tanabe, Hiroshi, Kanya Suzuki, and Yutaka Kaneko, Library Material Cataloging Rules and Exercises, edited by Nakamura Hatsuo and Maejima Shigekata. Tokyo: Jusunbo, 1981, p. 87. (Text in Japanese.) The ratio of Asian- to Western-language materials in Japanese libraries is 2:1; and the Japanese publishing industry ranks fourth, behind that of the U. S. A., Soviet Russia, and West Germany.

UNESCO, Statistical Yearbook, 1973-1985 editions. Calculating from figures given in the Yearbooks, the growth rates for publications during the period 1973-1983 are: for China, 12,493 titles

in 1978 to 31,602 titles in 1983 (53% growth); for Japan, 35,857 titles in 1973 to 42,977 titles in 1982 (20% growth); and for Korea, 7,396 titles in 1973 to 35,512 titles in 1983 (380% growth).

3. For some time the librarians of China, Korea, and Japan have expressed the need for such standardization and cooperation to promote mutual information exchange. The Conference of Directors of National libraries in Asia and Oceania, held every year, is one of their many commitments to exchange views and opinions to achieve such goals.

ROC: The National Central Library Newsletter, published in the Republic of China, reports various activities to promote national, regional, and international bibliographic organization and control, including workshops, conferences, and exchange programs between the Asian and Western worlds. For further discussion, see Donald V. Black and Margaret C. Fung, "Information Systems and Services in China and Japan." In Annual Review of Information Science and Technology, v. 18. White Plains, N. Y.: ASIS, 1983.

PRC: In the People's Republic of China, since 1949, many important changes have affected every institution and organization. In March 1971, Fang Yi, Vice-Premier of the State Council in Beijing, issued the "Online National Plan for the Development of Science and Technology." The nation is striving to meet the goals outlined in the Plan - to establish a number of documentation retrieval centers and databases, and build a preliminary nationwide computer network of scientific and technical information documentation retrieval centers. (See Black and Fung, "Information Systems.")

Korea: The Central National Library of Korea has been very active in computerizing the nation's bibliographic records. Following the guidelines of its "Ten-year Plan" (1978-1987) to achieve this goal, the library adopted LCMARC in 1980 and implemented KORMARC in 1981 by integrating, a la ISBD, the revised Korean Cataloguing Rules (3rd ed.), and quickly gained a number of supporters for the system from both public and academic libraries. (See The Information Book of the Central National Library; and Guidance to Printed Card of Korea - Based on KORMARC System. Central National Library, 1983.)

The Korean government's vigorous "Five-year Science and Technology Promotion Program" resulted in the merger of two important organizations - the Korea Institute of Science and Technology (KIST, opened in 1966) and the Korean Advanced Institute of Science

(opened in 1970) - into a coordinated organization, the Korean Advanced Institute of Science and Technology (KAIST) in 1981. With the integration of R & D activities and academic programs, KAIST will play an even more important role in the advancement of scientific knowledge and technological expertise in Korea. Communication and information technology have been two of its important priorities. (See Facts about Korea, 17th ed. Seoul: Samwha Printing, 1983, p. 112-113.)

Japan: By 1982 in Japan there were ten domestic online services using Kanji (Chinese) characters. Nakayama and Oikawa describe the development and consequences of special computer-processing systems at the University of Tsukuba - e.g., a special laser printer that provides high quality Kanji and Kana and roman letters. Nakamoto, et al., present a detailed description of various efforts to provide systems that can translate into Japanese materials in Western languages as well as in Chinese and Korean; Takahashi presents a logical proposal to develop and standardize an "Asian" character-set in order to encourage information interchange. (See Black and Fung, "Information Systems.")

Also see Proceedings of ICTP '83 (International Conference on Text Processing with a Large Character Set). Tokyo: Chinese Language Computer Society (CLCS) and Information Processing Society of Japan (IPSJ), 1983, 447p. The Conference was held to allow leading Asian researchers and engineers to discuss problems and share accomplishments in automating Asian characters - Chinese, Korean, and Japanese - for document-processing and international communication.

4. Spaulding, C. Sumner. "ISBD, Its Origin, Rationale and Implications." Library Journal, v. 98, p. 121-123, Jan. 15, 1973.
5. Includes UKMARC (BNB MARC, United Kingdom), USMARC (United States), DEUTSCHE MARC (Germany), INTERMARC (France), CANMARC (Canada), MALMARC (Malaysia), KORMARC (Korea), TAIMARC (Taiwan), JAPAN MARC (Japan), etc.
6. Chinese Cataloging Rules Working Group, Library Automation Planning Committee. Chung Kuo Pien Mu Keui Tse [Chinese Cataloging Rules] Taipei, Taiwan: National Central Library, 1982, 60p. (Text in Chinese)



Korean Cataloguing Rules for Description, 3rd ed. Edited by Jai Chul Lee. Seoul: Korean Library Association, 1983, 101p. (Text in Korean.)

Nippon Cataloging Rules, 3rd ed. Edited by Japanese Library Association, Cataloging Committee. Tokyo: NLA, 1977. 104p. (Text in Japanese.)

7. American Libraries, v. 16, no. 9, p. 617, Oct., 1985. The trans-border data-flow program, based at the National Library of Canada, is in the midst of a two-year factfinding period in which it will identify and quantify relevant patterns of electronic information flow, describe problems, and propose "realistic, workable solutions."
8. None of IFLA's meetings (in 1969, 1971, 1972, 1973, and 1974), held with the purpose of planning and implementing ISBD/UNIMARC, included Asian issues on the agenda or invited Asian representatives to planning groups. This, perhaps, indicates that IFLA's ISBD/UNIMARC was undertaken without any consideration of Asian countries or Asian materials. The terms "international bibliographic control" and "Western-world bibliographic control" were one and the same then, and are still used interchangeably.
9. "International Standard Bibliographic Description for single-volume and multi-volume monographic publications," recommended by the Working Group on the International Standard Bibliographic Description set up at the International Meeting of Cataloging Experts, Copenhagen, 1969. London: IFLA Committee on Cataloging, 1971, 36p.
10. Swanson, Gerald L. "ISBD, Standard or Secret?" Library Journal, v. 98, p. 124-130, Jan. 19, 1973.
11. Hegarty, Kevin. "Myths of Library Automation," Library Journal, v. 110, no. 6, p. 43-49, Oct. 1, 1985.
12. Diacritical marks applied in RLIN's romanization scheme for the Korean language:



Breve: "ˇ" Example: kun (pronounced as koon).  
 kūn (pronounced as kun)

Aspiration: "ʼ" Example: chin (pronounced similarly to jin)  
 chʼin (pronounced similarly to chin).

Apostrophe: "ʻ" Example: hanguk (can be pronounced as hang uk)  
 hanʻguk (pronounced correctly as  
 two characters).

Similar marks are employed by other romanization schemes for Japanese, Korean, and Chinese.

13. Swanson, "Standard," p. 1259. (The statement was made by Paul Fasana, of the New York Public Library, in an interview with Swanson, Nov. 9, 1972.)
14. Hegarty, "Myths," p. 48.
15. Anderson, Dorothy. "An International Framework for National Bibliographic Development: Achievement and Challenge," Library Resources and Technical Services, v. 30, no. 1, p. 13-22, Jan./Mar., 1986.
16. All three countries use an adapted version of the DDC/UDC Classification System - KDC for Korean, NDC for Japanese, and CDC for Chinese libraries. A list of subjects as they appear in the DDC or their class numbers are used in lieu of formally standardized subject headings such as LC's Subject Headings. (Exception: A library in Beijing uses the LC Classification System, but I do not know if this library uses LC Subject Headings. ROC's National Central Library uses modified LC Subject Headings.)
17. Dr. Gary Wang, President of Transtech International Corporation, invented the system through his Chinese terminal, called "Sinoterm."
18. Dr. Christopher S. George, Director of Research, Institute for Advanced Studies of World Religions at the State University of New York at Stony Brook, New York, invented the system.
19. Chu, Kuang-Fu. "Automation of Chinese Characters for the 80's," Chinese-American Librarians Association Newsletter, no. 26, Jan. 1982, p. 16-17.

20. Wang, Andrew H. "OCLC CJK Automated Library Information Network." Paper presented at the 38th Annual Meeting of the Association of Asian Studies, Chicago, March 23, 1986, p. 3.
21. Chu, "Automation," p. 17.
22. Black and Fung, "Information Systems." This article offers an excellent, extensive survey of information systems, products, and services, including computerization in China and Japan. The writer regrets that Korea was not included in the survey.
23. Library of Congress. Cataloging Service Bulletin (CSB), no. 118, summer, 1976; also, "Korean Romanization," CSB, no. 14, fall, 1981; "Japanese Romanization," CSB, no. 20, spring, 1983; "Classical Mongolian Romanization," CSB, no. 24, spring, 1984; etc.

Anglo-American Cataloging Rules, 2nd ed. Edited by Michael Gorman and Paul Winkler. Chicago: ALA, 1978, p. 3.

24. The first recorded interchanges between Korea and Japan were in the fourth century. Between the fifth and sixth centuries the flow from Korea into Japan included the Chinese writing system, various arts and handicrafts, and agricultural methods. Thus, Korea not only provided Japan with numerous aspects of culture but also served as a pathway for the infusion of Chinese culture. (See Nippon: The Land and Its People, Tokyo: Nippon Steel Corp., 1984, p. 53.)

The Korean alphabet (or syllabary) was invented by King Sejong, the greatest monarch of the Yi dynasty, who reigned ca. 1443 A.D. Since then both Han'gul and Hancha are used in the Korean writing system. (See Handbook of Korea, 3rd ed., p. 64-65.)

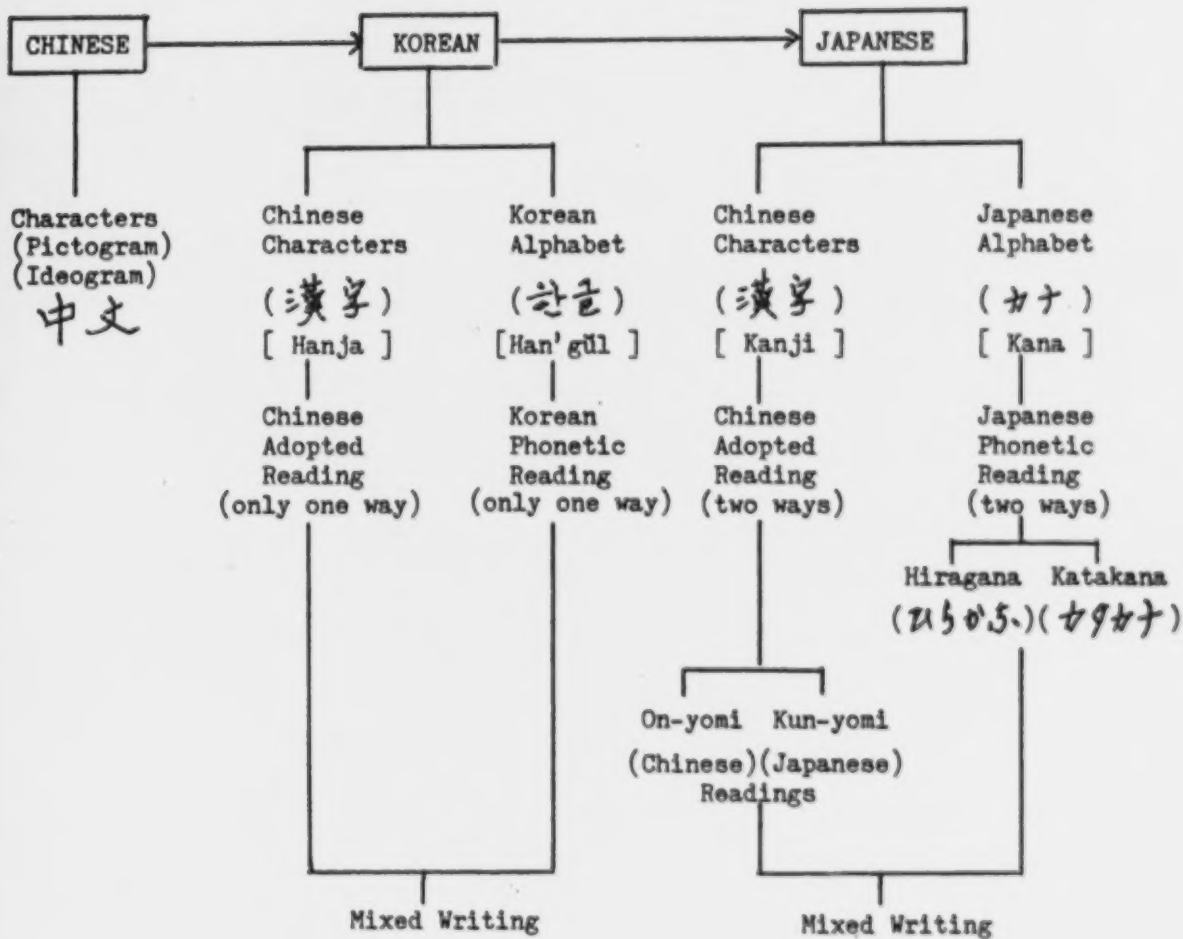
Japan also invented its own alphabet (or syllabary) - the Kana system (Hiragana and Katakana) - and uses all three systems in its writing. (See Nippon: ...)

25. O'Neill, Patrick Geoffrey. Japanese Names: A Comprehensive Index by Characters and Reading. New York & Tokyo: Weatherhill, 1972, p. 15, 98.

26. For the same reason that Westerners use romanization, Korea applies its own transliteration scheme, Koreanization (the application of Korean phonetic sounds to Chinese characters and Japanese Kanji), so that all Asian materials can be arranged according to the order of the Korean alphabet (Han'gul). Japan applies its own transliteration scheme, Japanization (the application of Japanese phonetic sounds to Chinese characters and Korean Hancha), so that all Asian materials can be arranged according to the order of the Japanese alphabet (Kana). (See Appendix D.)
27. Wang, Chi, and Robert Dunn, "Chinese Characters Computer Encoding Methods and Library Automation: A Bibliographical Survey" (Appendix), LC Information Bulletin, p. 95-100, March 26, 1982.
28. Dyer, Esther. "The International Situation: IFLA Meets in Munich," Wilson Library Bulletin, v. 58, no. 3, p. 190-193, Nov., 1983.

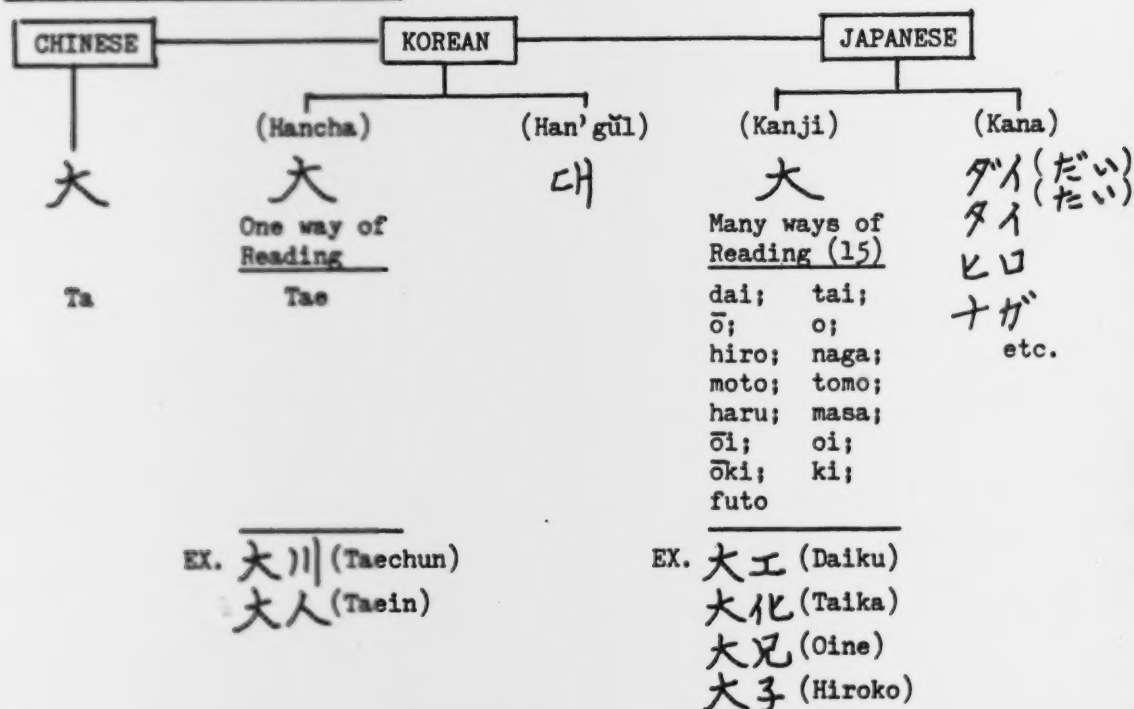
Jacques Tocatlian (UNESCO) emphasized the need for greater involvement on the part of developing countries in the affairs of IFLA. Currency restrictions, economic instability, and inconsistent support for delegates to attend IFLA meetings were cited as major barriers to such involvement.

"Developing nations want voice in IFLA Core Programs," in "Sleep no more at IFLA: Report on the IFLA 51st Council and General Conference, August 18-24, Chicago, USA," American Libraries, v. 16, no. 9, Oct., 1985, p. 617-618.

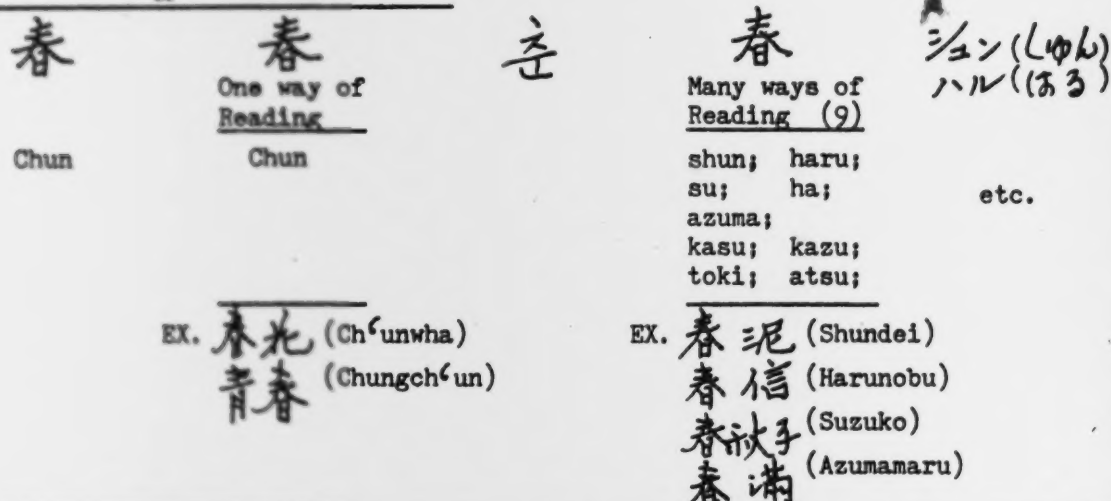
CHINESE, KOREAN, AND JAPANESE WRITING SYSTEMS

CHINESE, KOREAN, AND JAPANESE READING SYSTEMS

EXAMPLE 1: 大 = Big; Large



EXAMPLE 2: 春 = Spring (Season)



Source: O'Neill, Patrick Geoffrey. Japanese Names: A Comprehensive Index by Characters and Reading. New York & Tokyo: Weatherhill, 1972, pages 15 and 98.



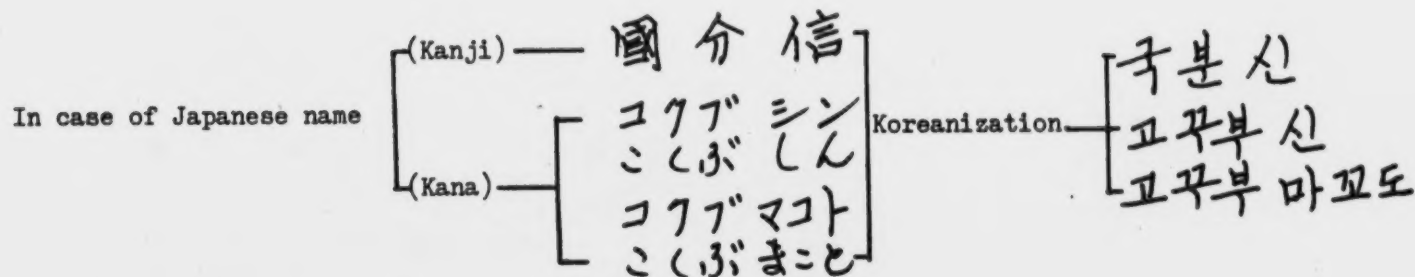
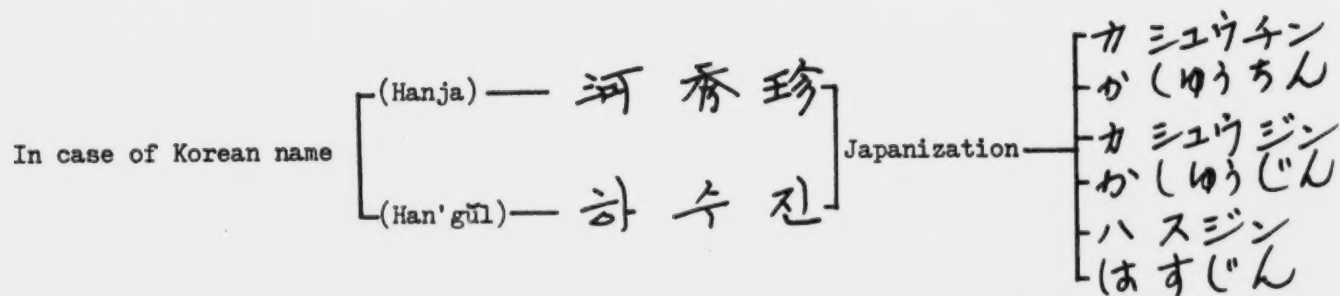
# ROMANIZATION SCHEMES APPLIED TO ASIAN SCRIPTS

(In case of Japanese author)

|                  |                     |         |                     |                     |
|------------------|---------------------|---------|---------------------|---------------------|
| 田辺 玄<br>(AUTHOR) | KOREAN LIBRARIES:   | 전 변 광   | (McCune-Reischauer) | 1) Chônbyon Kwang   |
|                  |                     |         | (M.O.E., 1959)      | 2) Chôn Pyon Kwang  |
|                  |                     |         | (M.O.E., 1983)      | 3) Jeonbyeon Kwang  |
|                  |                     |         |                     | 4) Jeon Byeon Kwang |
|                  |                     |         |                     | 5) Chonbyon Kwang   |
|                  |                     |         |                     | 6) Chon Pyon Kwang  |
|                  | JAPANESE LIBRARIES: | タナベ ヒロシ | KUN-YOMI (Hepburn)  | 7) Tanabe Hiroshi   |
|                  |                     |         | (Nippon)            | 8) Tanabe Hiroshi   |
|                  |                     |         | ON-YOMI             | 9) Den Hen Kō       |
|                  | CHINESE LIBRARIES:  | 田辺 玄    | (Wade Giles)        | 10) Tien Pien Kuang |
|                  |                     |         | (Pin Yin)           | 11) Dian Bian Guang |
|                  |                     |         | (ROC)               | 12) Dien Bien Guang |

# KOREANIZATION & JAPANIZATION

## Possible Variations





1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

62-MCULTP-2-E  
INF

NON-ROMAN AND MULTI-SCRIPT BIBLIOGRAPHIC DATABASES:  
BASIC ISSUES IN DESIGN AND IMPLEMENTATION

By

Alan M. Tucker

Associate Director for Library Systems Development

The Research Libraries Group, Inc.

Stanford, California, USA

KEYWORD(S):

Non-Roman  
Multi-Script  
Bibliographic  
Databases

For internal use only:

To be brought to meeting no:

Thursday 21 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

なぜ、非ローマ文字データベースに対して、思い悩まねばならないのか？ なぜ、単一のデータベース中に、二種類以上の非ローマ文字を結合させて、事柄を更に複雑にさせなければならないのか？ その第一の理由は、ローマ字化する方法の多様性と、既存の方法にも多くの例外的な適用があるからである。更にこの事に精通した利用者でさえ、原資料の表記を翻字した書誌記述の中に含まれる情報では、必要とする情報をすべて含んでいるわけではない、と断言するからである。今日の技術を最大限に活用することによって、非ローマ文字書誌データベースは、利用者にとって必要となる情報を提供することができるのである。

第二の理由は、二種類以上の非ローマ文字を含むデータベースだからといって、一つの非ローマ字のための一つのデータベースより、データベースの構築、維持あるいは利用が、本質的に複雑になるというわけではないからである。しかし、そこに存在する単純さを明らかにするためには、誰かがこの問題に対して実行してみなければならない。もし、何らかの原理を適用するとなれば、システムは、異なるレコードであるとか、異なる文字表記を含むレコード中のフィールドであるとか、あるいは、異なる文字表記を含むフィールド中の単語であることを、識別でき、そのデータを検索し、ソートし、表示することなどを適切に処理できなければならない。

現在のUS MARCフォーマットを構成する基本原理では、レコード、フィールドおよびデータ・レベルのそれぞれにおいて採用している文字セットの種類を指示するべきなのである。これによって個々のレコードは、文字セットの点では、自己規定能力を持つことになる。この原理は、レコードを作成するシステムにとっては、基本的なことであり、他の書誌システムとレコードを交換する場合にはなおさら必要なことである。このように、レコード自身で特定化されるので、システムに対して処理不可能なレコードは除外させることができる。即ち、処理できないフィールドをフィルターにかけるとか、無視させるとかをする。そして利用者に対しては、非ローマ字データを扱える利用者の端末機の機能に応じてデータを選択的に表示させられる。

異なるシステムの間で機械可読の書誌データを交換する場合には、もう一つの原理を尊重することが必要となる。すなわち非ローマ文字セットの標準化である。いくつかの文字は、現在、国際標準として認められている。その他は、国際標準に向けて処理検討段階にあり、更にその他に数多くあるわけだが一特定の国あるいは特定の言語圏内では、実質的に標準になっている。書誌記述および書誌調整のための非ローマ文字セットの開発における重要な課題には、歴史的な正確さや、中国語のように年代や地域によって文字の形の異なる文字そのものを処理できる能力、あるいはアラビア語、デーバナーガリー(インド文字)のように、正字法のある形式を処理できる能力である。文字コードの設計では、ソートする際のことを考慮しなければならないし、そして簡単な解決方法が常に可能であるというわけではない。

非ローマ文字による探索方法は、もし、データの自己規定性原理によって探索方法あるいは索引付がなされているのであれば、ローマ字形式での方法よりも多くの利点があり、そしてそれは、頭で考えている程複雑ではない。事実、簡単なユーザー・インターフェース、簡単なシステム・メンテナンスそして拡張性の準備があれば、多種の文字セットからエンターしたデータを一つの索引に結合することは可能である。

パーソナル・コンピュータの出現によって、アルファベット文字あるいは中国語、日本語、韓国語のように、連字活字印刷するような非ローマ文字(ロゴグラフ)の入力は、システムを開発する側の人間にとっては、非常に簡単になってきた。端末機上に表示される文字は、必ずしも期待どうりの見た目にも美しく喜ばれるとは、断言できないが、その改善のスピードはめざましく、特にレーザー・プリンターを使用した場合のプリンターからの出力は、そのまま出版できる品質にまで到達している。

## ABSTRACT

### Non-Roman and Multi-Script Bibliographic Databases

Bibliographic databases containing non-roman descriptions of non-roman bibliographic materials provide the highest and most precise information content for the user wishing to gain access to those materials. The technology is at hand to implement such databases, and thus to overcome the deficiencies of transcription and transliteration.

The paper examines specific issues which must be addressed in the design and implementation of multi-script databases, including the representation of non-roman scripts in machine-readable form, character set standards and use, searching using non-roman search arguments, the problems of sorting non-roman data, input and output devices and strategies, and systems with mixed technical capabilities.

## Introduction

Let us begin with two questions. First, why bother with non-roman script databases? And second, why, having gone to the trouble of developing the capability of handling two or more such things, apparently complicate life even further by combining them into a single database?

For many people involved in information science and bibliographic control, particularly but not exclusively in western Europe and North America, the first question is not nearly as frivolous as it might, at first glance, appear. Over the last twenty years or so, enormous amounts of money, time and skill have gone into efforts -- not all of them successful, of course -- to apply data processing technology to library practices, initially on a small scale in individual institutions, today on a massive scale involving international networks, and tomorrow in significantly changed forms involving the widespread linkage of heterogeneous systems. And at many points in those two decades, compromises were required on matters of detail if larger goals were to be achieved.

As an example, computers had been in use in libraries for a number of years before printers became available which could handle both upper and lower case roman letters. This aesthetic shortcoming was accepted because a more significant requirement was still satisfied by the system: it provided the necessary information to its users in a manner commensurate with the



technology of the day. It can be argued that in libraries accustomed to partial or complete romanization of non-roman bibliographic data, or in libraries so pressed by the need to process the incoming tide of publications that romanization could be viewed as the means to integrating the processing of non-roman materials into an otherwise already automated system, a database containing only records in roman characters does no disservice either to library staff or library users. In fact, the argument is regularly made -- and in some cases quite validly -- that some library processes, such as interlibrary loan, simply would not work if the computer provided references to non-roman materials only in the original script.

From one perspective, then, the argument is the same today for interlibrary loan as it was years ago for all-upper-case printing: the users are getting the information they need. From another perspective, however, the arguments are not at all the same. For it is perfectly clear that the ambiguities of romanization, the multiplicity of romanization schemes and the irregular application of any given scheme by even a relatively sophisticated user population combine to guarantee that the user is not being provided with all the necessary information contained in a bibliographic description if that description is fully romanized from another script. To quote from Hans Wellisch (who else, given the topic at hand?) on a specific, major issue: "... script conversion, and in particular romanization, is stretched beyond the limits of its usefulness when it is applied

as a bibliographic control tool, a task it is inherently incapable of fulfilling effectively." [1]

If we accept the premise that a bibliographic description in the language and script of the material described has the highest and most precise information content for the user who wants to consult that material (while not forgetting that there may be other needs and constraints which require different approaches -- I will return to this later), then given the requisite technical resources, the answer to my first question is the same as that I have used before: a non-roman database provides users with the information they need, in a manner commensurate with the technology of the day.

My second question can be more briefly answered, though the answer may be less immediately obvious. A database containing records in more than one non-roman script is not inherently more complex to build, maintain and use than is a single database for, say, Arabic, or several databases segregated by script. In my experience, this proposition is frequently met with an instantaneous, powerful, negative response. The difficulty is that one has to get very close to the problem before the underlying simplicity becomes apparent, and in looking at some detailed aspects of multi-script databases, that is precisely what we will now do.

-----  
[1] Wellisch, Hans. The Conversion of Scripts -- Its Nature, History and Utilization. New York: Wiley, 1978.

(I am sure that some readers are aware that there's a bit more to it than this in terms of internal processing. I have also restricted my examples to structured indexes, deliberately avoiding the issue of free-text searching. The use of embedded character set delimiters for the recognition of code-domain boundaries is crucial to free-text searching in a multi-script database and in multi-script records; the elaboration is more complex, but many of the underlying considerations are the same.)

Sorting non-roman data has some of the same problems and solutions as searching; indeed, in most implementations the two are closely linked, since efficient retrieval relies heavily on ordered index entries. But there are other, major problems, particularly when we look at what sorting means in the case of the logographic scripts.

Coded alphabetic character sets have generally been created with explicit ordering of the characters in mind: in ASCII, for example, the numeric value of the letter "a" is less than that of "b" which is, in turn, less than that of "c" and so on; the numeric values have an ordinal relationship which corresponds more or less reasonably to some real-world conception of the ordering of the graphics themselves. (I say more or less because there are inevitably some loose ends when one character set is used for many languages; thus, in Swedish, while "a" is the first letter of the alphabet, "a" with an angstrom above it, which the

### Data, Records, Characters

All the data in a database, roman and non-roman, textual and numeric, can be viewed as strings of bytes, a byte being the smallest directly-addressable unit of computer storage in most systems, having the capability of holding any of 256 discrete values and, more relevant to our concerns, the amount of storage occupied by a single character in the roman alphabet. We interpret those strings of bytes by imposing patterns upon them, thereby giving them logical and, at times, physical structure which corresponds to some meaningful entity, such as the bibliographic description of a book. The most widely familiar such structure for a library audience is the MARC record.

An automated system which handles MARC records (or, more likely, some derivative of MARC records, since the structure of MARC does not lend itself readily to online processing) typically deals with more than one form of data representation. Commonly, MARC records distributed by a national bibliographic agency, such as the Library of Congress or the British Library, are encoded in ASCII -- the American Standard Code for Information Interchange -- while the records as used in processing systems may be converted from ASCII to EBCDIC (the Extended Binary Coded Decimal Interchange Code), the "native language" of many computers. The "meaning" of a one-byte value in ASCII is not the same as that same value in EBCDIC; for example, in ASCII the value 107 represents the lower-case letter "k", while in EBCDIC

107 is the code for a comma. For most data, computer programs cannot readily figure out which coding system is present. In a given processing environment, certain assumptions are made, for instance, that records in the database will be in EBCDIC (I'm obviously thinking about an IBM-like environment), and that records exchanged with other computer systems, whether by magnetic tape or a telecommunications link, will be in ASCII.

Those global assumptions about the internal representation of data in a particular file or database are unlikely to have an equivalent validity in a database containing non-roman data, even though the ASCII-EBCDIC relationship provides a parallel model. For those non-roman alphabets in which each character can be held in a single byte -- as for the roman alphabet -- a given value, say 107 again, might mean this Russian letter or that Arabic letter or a Greek diacritical mark. Even in a database supporting only one non-roman script there is the possibility and, in those parts of the world where the roman alphabet is dominant, the virtual certainty that non-roman data will co-exist in the database, in individual records in the database, and even in individual fields in those records (think of a parallel title in Hindi and English) with roman characters.

There needs, then, to be a mechanism for defining what a given coded value means in a particular context. If we stay with the MARC record as a basis for this examination, we find that there are, in fact, two contextual levels at which the processing



system may need to be told explicitly what the data "means" in relation to some coding scheme. These are the record level, and the field level.

Until just a few years ago, it would have been possible to process MARC records containing non-roman data -- the National Diet Library is a perfect example -- but not to exchange them with another agency in a standard format. Nor, leaving aside the protocols of MARC record exchange, could they have been readily integrated into another processing system which handled not only Japanese but Hebrew and Arabic as well. The reason is that there was no agreed method of having the record carry information about the character sets -- the different coding systems -- it contained. The USMARC formats (for bibliographic data, but not yet for authority data) now support this information, making it possible for a system to know, at the record level, all the character sets that will be found in that particular record. How this works is, for our discussions here, not important; it's really not even very interesting unless you are writing programs that have to deal with it, and in the system that I am associated with, we have finally decided that we don't want users to know anything about it! What is important -- and particularly so in the multi-script database environment -- is the principle of having records carry information about themselves that can be used by the database and by the applications programs that connect the user to the data.

What kind of use? To begin with, ignoring records -- or parts of records; we've yet to deal with field-level data -- which can't be processed at all, which are only partially usable, or which are usable by some but not all users because of technical constraints. In the first instance, a national bibliographic agency might begin to include Chinese records in its regular distribution service. At the moment, very few subscribers are likely to be able to process such records, and might choose to filter them out during the loading of new records, by examining the field that identifies the character sets present in each record. In the last case, one can envision dial-up users of an online bibliographic system not all having the display capability to deal either with any non-roman data at all, or with some specific character sets. A search of the database using, say, an English language subject heading system, such as LCSH or Precis, might well retrieve records containing scripts which the user's terminal could not display. The host system could do any of several things: send the retrieved records to the user without any sensitivity to the receiving terminal's capabilities (thus probably painting the screen with garbage!); tell the user that records had been retrieved that satisfied the search criterion but could not be transmitted to his terminal; or, perhaps, use part of the record in romanized form to provide a meaningful display.

That brings us back to the second case above, the partially-usable record, and the question of field-level



character set information. This information is of two kinds: that which identifies those fields in a record that contain non-roman data (we are assuming here, and throughout this discussion, that the default or native character set of the computing environment we are working in is roman); and that which explicitly delimits the boundaries between contiguous roman and non-roman characters.

Using the first kind, a system which had no non-roman capability, but which received from a national agency or bibliographic utility records containing both roman and non-roman data -- as is the case for records produced using the USMARC format -- could strip the non-roman fields out and then add the rest of the record to its local database. Most turnkey library systems presently on the market essentially require the user to do either this, or alternatively to ignore the non-roman fields in processing; either way, field-level indication of the presence of non-roman data makes these operational choices possible.

With the second kind of information, we have at last arrived at the point of manipulating the data itself. We can use a real and simple example: an English-Japanese/Japanese-English dictionary, with the title given in both languages in their appropriate scripts. Let's further assume a processing environment in which roman characters are supported in the usual way (EBCDIC or ASCII -- it makes no difference for this illustration), and Japanese Kanji and Kana are encoded using the Japanese Industrial Standard

(JIS-6226), in which each character is represented by not one but two bytes. If the system is called upon to display the title of our dictionary on the terminal screen, it cannot just blindly shove the entire string of bytes comprising the title through the terminal's character generator. Why not? Because as we saw with the example of the byte value 107, the same number has different meanings depending on what coding system is being used to interpret it. The value 107 could not only turn up in the English portion of the title (Kenkyusha, after all, publishes dictionaries), but in a two-byte unit representing a kanji or kana as well. The system needs a way to know, at all times, what character set it is to use, since, as we noted earlier, the byte strings that make up the data itself are of little help.

Probably the most common way of signaling these character set or script "boundaries" is by means of escape sequences. Whatever specific form they may take in any individual implementation, escape sequences are generally groups of two or three bytes which do a number of things. First, and always, they get the program's attention in a completely unambiguous way (the program has to be looking for them, of course; byte strings are rather passive). Second, they provide the program with information about why they want its attention. In the context of our dictionary title, the escape sequence might tell the program that the data immediately following the escape is encoded using JIS, and then a little further on in the byte string, announce that the data following this escape is in roman again. In the case of the "escape to

JIS", the sequence might also indicate -- should indicate, in fact, if international standards are being followed -- that the JIS data is represented by more than one byte per character; it won't indicate how many more than one -- standards for the use of escapes for this purpose are based on the expectation that users of such data recognize the internationally registered codes; thus, if I receive records containing JIS-encoded data, there should be a signal telling me that, and I am expected to know that JIS means two bytes per character. Using this information, the program can put roman characters on the screen by passing the roman byte string through the normal character generator, and put kanji and kana on the screen by whichever method is implemented, another character generator, an external table, etc. These decisions are made possible by the embedding of signals about the coding scheme in the very data itself.

#### Character Sets: The Question of Standards

I have referred several times in passing to specific standards and to standards in general, and have been using the term "character sets" where earlier I used "scripts." Let us put these things together. There is one absolutely overriding reason for having standards for the way different scripts are encoded for data processing: communication. (If we don't want to communicate with some people, we encrypt messages, and were doing so long before there was any conception of computers or telecommunications; all the same, if the encrypted message

reaches its intended destination, the person or computer at that end knows how it was originally encoded, and can therefore understand it.) When we establish an unique correspondence between each of the characters -- the graphemes -- of a particular script and a numerical value, of one or two or n bytes in length, we have the basis of a character set for the processing of the languages which use that script, and for communicating the results of that processing with others. The character set may be promulgated as a national standard (e.g., JIS), and also be submitted to the review and approval process of ISO for possible acceptance as an international standard.

What is the current state of standardization for non-roman character sets? Non-roman data processing is certainly not in its infancy any longer, but I would suggest that in a number of areas, character sets in particular, it is in its adolescent phase at best. (I am taking as "adulthood" a situation in which ISO standards existed and were in use for the major scripts of the modern world.) A small number of examples are illustrative.

For the Cyrillic script, as used for Slavic languages in the Cyrillic alphabet, there is an ISO standard based on a national standard of the USSR, and for Old Church Slavonic a draft standard was sent out for a vote late last year. There is an ISO standard for modern Greek as well.

Arabic is represented by two non-ISO character sets at present, developed by different organizations. Neither has been proposed

to ISO, and neither covers the needs of non-Arabic languages written in the Arabic script, such as Urdu which has both vowels and consonants not found in Arabic.

Hebrew is in the draft proposal stage, which means that extensive discussion is still in progress, and issues remain to be resolved before an ISO ballot can take place. Outstanding questions include differences between Hebrew and Yiddish orthography, the inclusion or exclusion of vowel points, and rare characters occasionally found in minor languages written in the Hebrew alphabet.

The situation of character sets for East Asian languages is especially complicated. The Chinese, Japanese and Korean languages are all written using Chinese characters, either exclusively or in combination with a native syllabary, yet there are today at least five -- and possibly more -- character sets for the computer representation of Chinese characters. There is one each from the People's Republic of China, the Republic of China, Japan and the United States, and at least one from Korea. And between no two of these is there complete compatibility! They differ in the number of characters encoded, the number of bytes per character, the encoding of identical characters, and the treatment of lexicographical relationships between characters. The PRC and Japanese codes are national standards, published by appropriate government agencies. The other three are de facto standards, maintained and published by various



bodies; the American code is now the responsibility of the Library of Congress.

The development and standardization of character sets for non-roman scripts requires the solution of complex problems, problems which have arisen from forces such as evolution, religion and politics, all of which have at various times had an impact on orthographic practices. An important issue if we look at the suitability of character sets for bibliographic description and control is their historical accuracy, i.e., do they provide the tools to describe the writings of three thousand years ago, three hundred years ago or even the earlier years of this century when, for example, Turkic languages were "Cyrillified"? And if they do not, are the necessary descriptive compromises acceptable? Where characters have varied over time, as in Chinese, how important is it to be able to manipulate relationships between variant forms, that is, do we want systems to be able to search on one form of a Chinese character and automatically retrieve all of its variants as well? Where characters vary in the context of other characters, as in Devanagari and Arabic, what do we actually encode in the character set -- all the forms, or just the basic form of each character? If the answer to that is just the basic form (which I think it is), some implementation issues are simplified, such as searching and sorting, while others are made significantly more complex, such as display.

## Searching and Sorting

We are agreed, I trust, that non-roman descriptions of non-roman bibliographic materials provide the fullest, most precise and unambiguous information in support of access to those materials, and that, given appropriate technology, multi-script databases can be developed to store and manipulate that information. To my mind, providing access to the non-roman records in our databases by means of non-roman search arguments is a natural adjunct to the storage and display of the data, yet, like integrated multi-script databases, it is a concept whose underlying simplicity is not immediately obvious.

Searching a database for records which satisfy particular criteria -- an author's name, for example -- is ultimately a matter of comparing strings of characters, strings of bytes. Very simply put, the system takes the value of the user's search argument, in this case the characters (in whatever script) that make up the author's name, and compares that string of bytes to other strings extracted from the author field(s) of records in the file and normally stored in some sort of index structure. Where the strings match, the records corresponding to the index entries are retrieved and presented to the user for examination. Anyone familiar with online information retrieval systems knows that there are many variations and subtleties that can be added to make retrieval more sophisticated, and that even a basic retrieval process is more complex than my description. But at

bottom, searching comes down to a comparison of strings.

On the surface, the problem would appear to be embodied by the byte value 107 example we used earlier. That is, it is perfectly conceivable that a string of bytes representing the name of the Sanskrit playwright Kalidasa might be identical to one representing the name of a Russian economist, a Japanese journalist or a French chef. It is highly improbable, however, that someone searching a database for works by or about any one of those four would be interested in the other three!

A simple comparison of those strings, therefore, would produce a "false drop" -- the retrieval of items which do not really satisfy the search criterion. There are a number of ways to avoid this, and we will look at two. A multi-script database system could have several indexes for each index category, one for each supported script. Thus, using the previous example, the system might have a roman personal name index, a Cyrillic personal name index, one for Devanagari and one for Japanese as well (though in the latter case it might be an index for all languages based on Chinese characters). We have seen that information about character sets can readily be embedded in a bibliographic record; a database system could make use of that information to segregate index entries explicitly. In the abstract this works perfectly well. One must assume that the user knows what character set he is using when entering the search. But it significantly complicates the user interface,

since the number of indexes the user may have to be aware of is the number capable of containing non-roman data multiplied by the number of character sets the system can handle. And as neither of those factors is necessarily static -- both new indexes and new character sets might be added -- maintenance of the system is further complicated too.

An implementation of the sort just described relies on embedded information identifying character sets. That information can also be used to simplify indexing and searching of non-roman data if it is maintained as part of both the search argument and the index entries derived from the bibliographic records. In other words, if each of our four personal name examples were preceded by a unique value indicating the character set -- if that value is treated as an integral part of both argument and index -- then a simple string match will not produce false drops, but will correctly discriminate between strings which have no logical relationship. If this approach works (and it does), the system can handle all personal names with a single index, and we are back at the simple level of straightforward string comparisons, because the unique character set identifiers logically segregate different character sets in the same way that the rules of alphabetization segregate the "a"s from the "b"s. The simplicity can almost be said to follow from ignoring the external problem, since, as I know from experience, trying to answer questions such as "How do we combine Cyrillic and Chinese with an existing roman indexing system?" can make system designers' heads spin.

language recognizes as a separate character, is the twenty-seventh. ASCII isn't much help.) The same principle is evident in the Cyrillic standard, and in the draft Hebrew standard which even takes account of the relative ordering of final and non-final forms in Hebrew alphabetization.

Sorting of such ordinally-structured character sets can therefore be accomplished using standard sort packages if the same principle described for searching is maintained -- treatment of the character set designators as integral to the data being sorted. (I am thinking here of what one might call data suitably prepared for sorting; the distinction is necessary as heretofore MARC records have been the general example, and data in a MARC record is definitely not suitable for sorting without preparation, a point to which I will presently return.) If we are sorting words, Cyrillic words will sort together, subarranged by the code values of the individual letters. Where they will sort relative to words encoded as roman characters or Greek letters could be left to the sort package as well, in which case the order would be that of the codes identifying the separate character sets. If that were unsuitable -- as I expect it would be in most implementations -- a canonical ordering of scripts could be achieved by pre-processing the data to replace the order-independent values of those identifiers with other values selected to achieve the desired sequence.

We normally understand what order means in alphabetic scripts, at



least in relation to a particular language if not globally, and that knowledge can be conveyed by the codes in a character set. But what is "order" in Chinese, in Japanese or in Korean? Is there a single order, as in the English alphabet, or several? Can we convey any order in the coded character set, and if so do we have to sacrifice any other information? If we cannot carry the order within the codes, can we sort such data at all? I will try to answer this small subset of the possible questions that could be asked on this topic.

First, there are a number of possible "orders." One simple one is by the total number of strokes in a character, such that a one-stroke character precedes a two-stroke character, and so on. But since the number of strokes in characters ranges from one to about twenty-eight, and there are no less than 50,000 Chinese characters, this method provides only a first order discrimination. A refinement on ordering based on strokes is to use the Chinese radicals -- every Chinese character contains one of 214 "radicals", themselves characters -- for initial sorting, followed by subarrangement by the remaining number of strokes in the character; this radical-stroke arrangement is used in many dictionaries. Given agreement on the sequencing of characters having identical radical/stroke counts -- a questionable proposition -- a code could represent the order of all the characters in the character set.

But existing Chinese character codes have been based on other

principles, such as frequency of use or the relation between variant forms of characters, and may even be layered so that "order" takes on a third dimension beyond the linearity of an alphabet. Another possible sequence, then, is phonetic, using either an accepted romanization scheme or a native phonetic representation, such as kana for Japanese, hangul for Korean, or the Chinese National Phonetic Alphabet. This approach, however, makes it impossible for the character code to carry order information, given existing codes at any rate, since the fixed-length two-byte JIS code, for example, cannot hold the number of bytes required to render a kanji phonetically in most instances (e.g., the Kanji meaning "new" is coded as two bytes; it can be expressed as two kana -- four bytes -- if it is pronounced "sin", or as five kana -- ten bytes -- if pronounced "atarasii." Even in the widely-used Hepburn romanization, its shortest phonetic equivalent takes three bytes!) And how is a system to know how to pronounce the character if there are different possibilities?

Perhaps the answer is, it can't. And perhaps the answer to an earlier question is, direct sorting of Chinese, Japanese and Korean data based on the character codes presently in use, is not possible. Another technique is required. One such technique is -- getting back to our MARC record -- to carry a phonetic representation in addition to the precise (e.g., Chinese characters, plus kana or hangul for Japanese or Korean) representation within the record. For Chinese and Japanese, such

data could be used as sort keys and would produce meaningful and consistent results; for hangul the "meaningfulness" is debatable.

A variation on this is to use supplementary information about the characters in a character set to generate sort keys as required, as part of the pre-processing that is virtually always necessary to make data in MARC records suitable for sorting. This pre-processing typically involves some or all of the following in the construction of a sort key which is sorted as a surrogate for the actual record: removal of control information, such as subfield delimiters; conversion of data from a mixture of upper- and lower-case to either all upper or all lower, in character sets having case differences; inclusion or exclusion of specific subfields; and the concatenation of fields not necessarily contiguous in the record. Were the data available in an associated database, radical/stroke and/or phonetic values could be substituted for the character codes in the sort key, which would then be passed to a standard sort package.

#### Input and Output

When I began working with the computer processing of non-roman bibliographic data about eight years ago (I started with Chinese, incidentally, not wanting to do anything that was too easy...), some of the most pressing concerns had to do with input devices. For alphabetic or syllabic scripts, many solutions had existed for years: there is nothing new or exciting about a Russian

typewriter, which translates directly into the keyboard of a Russian terminal which, with minor modification, becomes capable of handling an extended Cyrillic character set.

The interesting problems have to do with scripts which either run in opposite directions (as Hebrew and Arabic do when used in conjunction with roman or Cyrillic -- or, of course, vice versa if your daily context is Hebrew or Arabic), or whose characters change form depending on adjacent characters, such as Arabic and Devanagari. In each of these cases, in order to operate properly for continuous typing and for the insertion and deletion of characters or character strings, the keyboard device needs to be context-sensitive; that is, it needs to know not only where the cursor is -- the position at which a character will be entered the next time a key is struck -- but also what is on either side of the cursor, what character set is currently invoked and what direction the cursor is moving, right or left, up or down.

When the only tool available was a dumb terminal, this kind of problem could be solved crudely if at all. Early intelligent terminals were more flexible, and allowed the manufacturer or system integrator to begin to address the context issue. But the real breakthrough, as with so much that we see in user-oriented computing, was the personal computer, which now provides sufficient power and flexibility to make possible sophisticated, user-friendly solutions to the problems outlined above. Today one sees very few systems for the input -- and, in many cases,

manipulation -- of non-roman data, ranging from the relative simplicity of Cyrillic to the complexity of Japanese to the special problems of mixed roman and Hebrew, which are not implemented using some kind of PC technology. And the real thrust of most design efforts is not simply, as before, getting the data in, but getting it in in a way that appears as natural as possible to the user. Part of this has to do with keyboard layouts, and part with the logic and aesthetics of display.

For most alphabetic scripts, there have been typewriters for many years, and there are a multitude of real and de facto standards, and variations on themes (e.g., the layout of letters on Italian typewriters differs in two positions from that of English or American typewriters; and there are at least four layouts for the Korean hangul typewriter.) In most cases these layouts are directly adaptable for use on a computer input device.

But for logographic scripts the situation is quite different. Until very recently keyboards for typing or typesetting or electronic text composition were extremely cumbersome and, in general, not easily used by any but professional operators. With the beginning of interest in the more wide-spread application of Chinese-character data processing, considerable attention was paid to the problem of keyboard layouts and input methodologies. Some of the design goals, occasionally in conflict with each other, were: simplicity; ease and speed of operation; large character repertoire; precision, i.e., lack of ambiguity in the



selection of characters.

As a result of these efforts, there exist today literally dozens of methods of entering Chinese characters, and possibly as many as several hundred separate implementations of those methods in different systems. I won't go into detail on any of them, but will just note in passing that the methods fall generally into a few conceptual families: large keyboard systems in which complete characters are represented on a large number of keys (e.g., fourteen characters on each of 256 keys); composition systems, in which characters are "assembled" as the operator inputs components selected from a smaller keyboard, perhaps even a standard QWERTY keyboard; phonetic conversion systems, in which input in, say, kana or romanization is automatically converted to the correct logographic representation; tablet systems, in which the operator writes on a data tablet and the system interprets positional information to generate the correct characters; and finally, at a very early stage of development, voice recognition systems. The majority of operational systems today are PC-based.

The PC revolution has greatly improved the quality of output available for non-roman applications, too. The principal considerations in the design or selection of systems are essentially these three: resolution, speed and aesthetics. Resolution here really means two things -- how clearly can characters be represented, and how much data can be displayed on

a terminal screen using clearly represented characters. The first is a function of how many addressable points there are on the display device, and of how many of those points are used in the rectangular display cell within which each character appears. A 7x9 cell may be perfectly suitable for the characters you are reading right now, but does not seem adequate for Hebrew or Arabic, and is definitely unsatisfactory for Chinese, where even a 16x16 dot matrix produces characters which, though legible, have no claim to typographical quality. But at some point a balance must be struck. If we go back to our example of bibliographic records, the size of characters and the number of characters per screen have to be considered in relation to the amount of information -- not characters -- we want to be able to display at one time for a particular purpose. Chinese characters created using a 128x128 dot matrix are quite beautiful, but a fairly brief bibliographic description such as that on a typical catalog card would not fit on a single screen.

The "beauty" of characters is an issue that will become more and more important as higher resolution graphics make it possible to use more display points in less space, that is, to display more detailed characters at a smaller scale. Today there is still a need for compromise in this area; in some ways it is reminiscent of the days of all-upper-case printing that I referred to earlier. Hard copy output devices are well ahead of displays in this regard, with many laser printers now available capable of producing extremely finely shaped, well resolved characters in an

un-exaggerated scale.

The issue of display speed arises largely because many non-roman display systems operate in graphics mode, drawing each character point by point. The impact of the microprocessor's speed can be readily seen when using the system to edit a full screen of text, when the insertion or deletion of a block of data can necessitate the re-drawing of a large area of the screen. Depending on the application and the character set being used, the relative speed of, for example, an IBM PC and an IBM PC/AT may be both perceptible and annoying.

(There are, incidentally, some very interesting programming problems associated with displaying non-roman data, particularly when a mixture of right and left cursor movement is involved, or when the data contains both single- and multi-byte character representations. Unfortunately there is no time to explore those here.)

#### Environmental considerations

Under that rather unusual title, I would like to conclude by talking about some important general issues affecting the design and use of non-roman and multi-script databases for bibliographic data. The environment I am considering is the one in which a system is going to be used -- the people who use it, the dominant language and script of those users (if there is just one), and the rules of bibliographic practice used to give the database

logical coherence. And despite the instructions to participants here to be "general in nature" and "neutral in tone," I will use my own particular experience to illustrate these concerns and to leave you with questions to be answered in your own environments.

The first issue has to do with the general utility of data in a heterogeneous systems environment, one in which the technical capability to deal with non-roman data is not universally available to users of a system. In the United States at this moment, only a relative handful of institutions can process Cyrillic or Chinese bibliographic data in a network environment. Yet the information contained in those Cyrillic and Chinese records is of potential value to people who have access to the network, want access to the materials, but can't display the non-roman data on their terminals. The solution we adopted was to modify the MARC record structure to provide for parallel fields -- for example, one field for the non-roman title, another, linked to the first, for the romanized title. And in a particular implementation, we required that certain of those parallel fields always had to be used, so that any network user, at any terminal, could make sense of the records, and so that other systems with which we communicate data could add usable data to their own files. If you are acquiring or developing a system to handle non-roman bibliographic data, and if you find yourself in an environment in any way similar to the one I've described, how will you deal with this problem?

A second issue arose when we were beginning to design our Hebrew implementation. Roughly stated, the question is: we -- in the United States -- live in an environment in which the roman alphabet is dominant, that is, a left-to-right world; how, then, do we accommodate the apparently conflicting factors involved in handling MARC records (with their own rules for the direction of data flow), and the mixture of roman and Hebrew characters which will naturally occur in any record produced in accordance with AACR2? Our answer, not arrived at easily, was to say, "Yes -- we live in a left-to-right world, and our design will necessarily reflect that fact." The effect of that decision is that, if we look at a display of a MARC record including all its tags, indicators, etc., the tag for a Hebrew title will not be at the right edge of the screen, it will be at the left; and though the data in each subfield will read from right to left, the subfields will be sequenced from left to right. (This is for an active display, of course, one into which a cataloger is entering and editing data; we recognize that passive displays, meant to provide information to a reader, have to be handled differently and have designed the system accordingly.) A system in Israel would not look the same as ours. If this problem is significant in your environment, how will you solve it?

I will leave you with those questions, and with any others which this discussion may have raised. We are at a very early stage in the development of multi-script bibliographic databases. Those of you who have developed or used such systems know both the



excitement and the frustration of pioneers everywhere. Those of you who are now about to get involved can benefit from the successes -- and mistakes -- of those pioneers.



1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTIScript  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

63-MCULTP-3-E  
INF

Computer Processing of Southeast  
Asian Scripts, with Particular Reference  
to Jawi and Thai

by

Edward Lim Huck Tee  
Chief Librarian  
Universiti Sains Malaysia, Penang, Malaysia

KEYWORD(S):

For internal use only:

To be brought to meeting no:

Thursday 21 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

序：東南アジア言語は方言を含めると数百にも及ぶ。ごくわずかに非ローマ字を使用する言語・方言があるが、過去にキリスト教伝道会によってローマ字化がすすめられ、公用語として採用されてきた。東南アジア言語の文字処理において、非ローマ文字の存在は自動化の障害となっている。本稿では、SEAPRINTプロジェクトにおけるローマ字とThai文字の処理の経験ならびにマレーシア工科大学で実施されようとしているJawi文字の処理について報告する。

文字生成の原理：非ローマ文字をサポートできるコンピュータ・システムを作るには、特別なキーボードの他に、特殊なROMキャラクタ・ジェネレータを考案しなければならない。プリンタを含めたこの種のハードウェアの改良の他に、適切なソフトウェアも必要となる。通常、非ローマ字を含む文字処理をするためには、以下の問題を解決しなければならない。

- \* 7ないし8ビットを使用する言語のマッピングの問題。この場合は、その言語が128ないし256のコードに簡単にマッピングされるかどうかが課題となる。中国語のように256のビット・パターンですべての文字を表現出来ないものもある。
- \* 言語固有の問題。この問題を解決できるソフトやハードは、必ずしも簡単に見つからない。
- \* ローマ文字と非ローマ文字のレコードを同時に処理するさいの問題。

Jawi文字：Jawiは、アラビア文字を適用し、いくつかの追加文字を持っている。Jawiは、35字のアルファベットとlam-alifとhamzahからなる。マッピングの問題はほぼないが、文字の語の中での位置関係によってアルファベットの形がわずかに異なり、結果的に100以上の異なる文字種がある。目録や書誌レコードでは、ローマ字とJawiの文字セットを併用する必要があるので、ASCIIやEBCDICコードを使って文字をマッピングする必要が生ずる。しかし、現在Jawiの文字セットを符号化する基準は用意されていない。Jawiの文字セットのコンピュータ処理を複雑にする以下のようないくつかの要因がある。

- \* 文字は右から左へ、数字は左から右へと書かれる。Jawiの文字セットをローマ字と同時に処理し、その際に読み方や書方が左から右であることが処理を複雑にしている。これまで2つの解決法が試行されてきた。第一は、画面分割ないしはウィンドーである。第二は、カーソルの移動を一方向のみにし、ローマ字からJawi文字へ移る際に切り換えられるようにする方法で、後者が優れている。
- \* Jawi文字は語の中での位置関係によって形が異なるので、システムに自動的に正しい形を選択させなければならない。そのために、文脈分析できる処理が必要である。
- \* ソーティングの問題－辞書体配列の基準がない。
- \* 特別な入出力装置の採用。
- \* 異なる文字セットを同定するソフトウェアの問題。

Thai文字：Thaiの書法は、44の子音、32の基本母音、声調符号、特殊記号と句読点から成り、マッピング上の問題はない。事実、タイIBMは、EBCDICを使う72のコードを持っている。Thaiの文字セットは、lower caseにローマ字と特殊記号をコード化しているので、upper caseでローマ字とThai文字を併用する。これはIBMを使用し、そのハードやソフトの設計をあまり変更させないようにしたための措置であった。

Thai文字のコンピュータ処理を複雑にする要因として以下があり、SEAPRINT Pilot Projectで解決した。

- \* いくつかの母音が子音の前に書かれ、子音の後の場合でもその上や下に書かれる。
- \* 声調符号が、すべて上付き文字である。
- \* 語間のスペースがほとんどない。
- \* 配列上の問題。

ソフトウェアの利用：3つのタイプのソフトが利用ないし試行されている。1)ワープロ、2)DBMS (MINISIS)、3)MALMARC (マレーシアMARC)

## ABSTRACT

There are several hundred Southeast Asian languages and dialects, but only a handful of them use non-roman scripts. In this paper, attention is focussed on the computer processing of Jawi and Thai. The problems discussed include:

- \* Problems relating to the mapping of the languages using 7-bit or 8-bit coding scheme.
- \* Problems that are peculiar to the languages concerned, and for which software and hardware solutions have to be found.
- \* Problems relating to the processing of records containing a mix of roman and non-roman scripts.

An outline of some available software which can solve some of the problems mentioned above is provided. The technology to process a wide variety of scripts is available. The problem is to persuade computer manufacturers to design systems for which there is a limited market.

LHT/nsnl/seascrpt.ab

COMPUTER PROCESSING OF SOUTHEAST ASIAN SCRIPTS  
WITH PARTICULAR REFERENCE TO JAWI AND THAI

by

Lim Huck Tee  
Chief Librarian  
Universiti Sains Malaysia

INTRODUCTION

It would be impossible for one person to deal with the computer processing of the wide variety of languages and scripts used in Southeast Asia, since the number of languages and dialects would number several hundred. As a result of the efforts of Christian missionaries, a large number of Southeast Asian languages make use of the roman alphabet and perhaps some special characters to indicate tones and sounds not normally found in European languages. Most of the nations of Southeast Asia have adopted the roman script as the official script of their languages, and these include the major Southeast Asian languages such as Tagalog (the official language of the Philippines), Bahasa Indonesia and Bahasa Malaysia. Nevertheless, the libraries in these countries have still to deal with publications in non-roman scripts. Thus the national bibliography of Malaysia also lists publications in Jawi, Chinese and Tamil; the Indonesian national bibliography has to cope with occasional publications in Javanese, while the Thai national bibliography lists almost exclusively publications written in the Thai script.

As long as the bibliographies and catalogues of Southeast Asian libraries were manually produced, there was no serious difficulty in producing catalogue cards or bibliographies using a wide variety of scripts, since these could be handwritten or produced using typewriters with specialised keyboards and type faces. However, computerisation has made it more difficult to handle non-roman script records because computers as designed by the developed countries are used in environments where the roman character set prevails.

With the exception of the Thai National Library which has developed a limited capability for processing Thai scripts, the national libraries of Malaysia and Singapore have tried to get around the problem by using transliteration schemes for non-roman scripts. Obviously, this is not a very satisfactory solution, and efforts are currently being made to acquire computer hardware and software which can process the materials using the original scripts. While the technology to process a wide variety of non-roman character sets is available, most of the current computer manufacturers do not think it worth their while to invest time and money in producing systems that can process non-roman Southeast Asian scripts.



In this paper, I will only deal with some of the problems and solutions which have been tried out or are being considered with respect to the processing of Jawi and Thai. My experience with processing Thai script came about in connection with the SEAPRINT pilot project - a project which attempted to produce a computerised regional bibliography of selected Southeast Asian imprints and involved the simultaneous processing of roman and Thai scripts. Work on Jawi script is being carried out at the Universiti Teknologi Malaysia (which is trying to produce a Jawi word processor), and by a graduate student at Universiti Sains Malaysia (who has attempted to program an EPROM chip to produce the Jawi character set on the monitor of an Apple computer).

#### PRINCIPLES OF CHARACTER GENERATION

A simplified explanation of how characters are generated on a monitor or produced by a printer is a useful preliminary step towards an understanding of the problems relating to the character generation of non-roman scripts. Computers make use of a binary system to communicate. Humans, on the other hand, need to communicate using words (in a script that they can recognise and understand). Data passing between the computer (or more accurately, the CPU - Central Processing Unit) and output devices such as a CRT (Cathode Ray Tube) or printer comprises groups of seven or eight bits with each group representing an alphanumeric or other character depending on the coding scheme used. The two most common coding schemes are ASCII (American Standard Code for Information Interchange) which uses a 7-bit code, and EBCDIC (Extended Binary Coded Decimal Interchange Code), an 8-bit code which is used primarily in IBM computers.

*Display using a monitor.* The basic peripheral for entering characters is the keyboard, and the basic peripheral for displaying them is the monitor which is built around a CRT. CRTs generally use what is known as raster scan techniques to display images on the screen. In a monitor designed for the display of alphanumeric characters, the images of whole characters are produced by a character generator ROM (Read Only Memory) chip. Data are fed into the system in the form of standard codes (e.g. ASCII or EBCDIC) and these are converted by the character generator into predetermined bit patterns for each determined matrix of dots, e.g. 7 x 7 or 7 x 9 bits. (Fig. 1 provides an example of a roman character set using a matrix of 7 x 7 dots, while Fig. 2 shows a proposed 7 x 13 matrix for Jawi characters).

The ROM character generator can be programmed with any set of characters, and it can be activated when an appropriate key on the keyboard is pressed. This in turn can cause an appropriate pattern of dots to be displayed on the monitor.

*Printers.* There is a wide variety of printers available. In general, one of the ways of classifying printers would be to differentiate between printers where the characters are already fully formed and are embossed in reverse, and printers where the characters are shaped by a combination of dots. Line and daisy

wheel printers are examples of the former, while dot matrix printers are examples of the latter. In the case of a printer where the character set is fully formed, the ASCII character code sent to the printer is translated into a motion that causes the desired character or characters to strike the ribbon creating a printed impression of the character(s) concerned on the paper. In the case of dot matrix printers, the ASCII code causes a ROM chip to send the appropriate pattern of dots to the printer.

To produce a computer system that would be capable of supporting non-roman character sets would require the design of special ROM character generators as well as special keyboards. For fully formed character printers, special print trains or daisy wheels would have to be designed; in the case of dot matrix printers, any shape can in fact be produced using specially written programs to generate the required graphics or by means of a pre-programmed ROM chip. This does not mean that the solution to the problems of processing non-roman script languages is purely a hardware one. Appropriate software is also required to deal with some of the special characteristics of the language(s) concerned. In general, solutions have also to be found for the following:

- \* Problems relating to the mapping of the languages concerned using a 7-bit or 8-bit coding scheme. This implies that the languages concerned can be easily mapped in 128 or 256 different codes. This is not always possible for some languages. For example, it is impossible to represent all Chinese characters in 256 combinations of bit patterns.
- \* Problems that are peculiar to the language being processed, and for which software and hardware solutions are not always easily found.
- \* Problems involved in processing records in roman and non-roman scripts simultaneously.

In this paper, I will examine the computer processing of Jawi and Thai scripts in terms of the problems mentioned above, and discuss some possible solutions to these problems.

#### JAWI SCRIPT

*"Mapping" problems.* Jawi is essentially an adaptation of the Arabic script, with a few additional characters not found in Arabic. It is thus an alphabetic language. There are thirty-five letters in the alphabet but of these one (lam-alif  $\text{لـ}$ ) is merely a combination of lam ( $\text{ل}$ ) and alif ( $\text{ا}$ ), and another hamzah ( $\text{ء}$ ) is an indication of pronunciation rather than a letter. In general then there are no problems in mapping Jawi to a 7-bit or 8-bit coding scheme. However, the alphabets have slightly different shapes depending on their position in a word, e.g. a character appearing in the beginning of a word may have a different shape from the same character appearing at the end or in the middle of a word. (See Fig. 3 for examples). Consequently,

there are more than 100 different shapes. Since it is also necessary to mix roman and Jawi character sets in a catalogue or bibliographic record, the script needs to be mapped using an extended ASCII code (which will allow up to 256 characters to be mapped) or an EBCDIC coding scheme since these would allow up to 256 different 8-bit combinations to be mapped. Unfortunately, however, no standard exists for the coding of the Jawi character set at the present time.

*Problems peculiar to the script.* While there may be no major problems in mapping the Jawi character set to an ASCII or EBCDIC coding scheme, the Jawi character set contains some unique features that complicate its computer processing. These characteristics include the following:

- \* Jawi characters are written from right to left, but the numerals are written from left to right. Processing becomes complicated when the Jawi character set is processed simultaneously with say a roman character set, where the reading and writing direction is from left to right. Two solutions have been tried. The first uses the concept of the split screen or window whereby the right half of the screen is reserved for the Jawi letters, and the left half for roman characters. The second solution (which is superior) ensures that the cursor moves in only one direction, and that when a switch is made from say, a roman character set to Jawi, the insert mode is turned on so that as the Jawi characters are entered, the characters move to the right. This solution has three advantages: since the roman characters are read from left to right and the Jawi characters from right to left, it is possible to mix characters on one line of text; there is no need to have special software to control cursor movements; and the carriage return and line feed can be used as for romanised text.
- \* As shown in Fig. 3, the Jawi characters may change their shape in accordance with their position in a word. A solution must therefore be found to have the computer system select automatically the correct form. This can be done using a process called contextual analysis, whereby each character as it is entered is analysed in terms of its context within a word, and the appropriate shape is then generated. Occasionally, it may be necessary to generate a character in its "isolated" form; consequently, the software must also allow the contextual analysis to be by-passed.
- \* Sorting is not a straight forward process. For example, as pointed out above, lam-alif is really two letters and should be sorted as such. Similarly, all the forms of the hamza should be sorted identically. Furthermore, there is no standard dictionary ordering of the characters as in the roman alphabet. The traditional scholar,

for instance, may place words in a dictionary in a different order from that of the modernist. Since the dictionary order may be different from the collating sequence as mapped using a coding scheme, there would be a need to write a special sorting program with special condition checks.

*Problems in processing mixed scripts.* The major difficulty in processing Jawi simultaneously with other scripts arises from the fact that Jawi is written in a different direction from most other scripts. There are also some problems which are inherent in most mixed script processing systems. These may be summarised as follows:

- \* Special input and output devices have to be employed to process the mixed scripts, and these devices must have a standard means of identifying the different character sets.

- \* When a catalogue or bibliographic record contains two or more character sets, the software should be able to identify the different character sets. But this is by no means easy. There are two approaches to solving this problem. The first approach is to assign unique codes to the characters of the different scripts. However, this has the disadvantage of limiting the number of character sets that can be processed simultaneously, because most computer manufacturers use either a 7-bit or 8-bit code to represent characters internally. The second approach is to prefix each character string with a character set indicator; this has the advantage of allowing many different character sets to be mixed within a record. There is, however, one caveat. Although it would be possible to mix several different character sets in one record, it is usually difficult to mix character sets within a field. The reasons are as follows:

- During the input and output of data, complicated processing is needed to differentiate between the various character sets used.

- The filing sequence between different character sets has to be arbitrarily determined because there is no recognisable sequence that can be adopted.

- Difficulties will be encountered if the mixed character sets have reading/writing directions which are different.

- If it becomes necessary to index a field so that various access points can be provided in an inverted file, problems arise because the list of inverted terms comprising different character sets cannot be sorted or ordered in a desired sequence.



The solution is of course not to allow mixed character sets within a field, but to place different character sets in different fields or in repeatable fields.

## THAI SCRIPT

*"Mapping" of Thai script.* The Thai writing system consists of 44 consonants, 32 basic vowels, tone marks, special symbols and punctuations. Thai is also a mappable script. IBM Thailand has in fact coded 72 of these characters using the EBCDIC. In order to allow the system to be used on standard IBM input and output devices, the Thai character set has been coded in the areas reserved for lower case roman alphabets and other special symbols. As a result, the Thai character set can only be mixed with upper case roman letters. The decision was probably a pragmatic one - to facilitate the use of existing IBM hardware without too many changes to the hardware and software design. (See Fig. 4)

*Special problems in processing Thai script.* The following characteristics of Thai complicate its computer processing:

- \* Some vowels are written before the consonant, some after the consonant, some above and some below.
- \* Tone marks are all superscripts and are placed above the right hand side of the initial consonant of a syllable; if two consonants are used together at the beginning of the syllable, the tone mark is written above and to the right of the second consonant. If there is a superscript vowel, the tone mark is placed above the vowel. As a result of this, the Thai character set occupies the equivalent of two physical lines of space. Thus, if the letter "A" for example occupies a matrix of 7 x 7 dots, then the Thai character with its superscript vowel and tone marks would require a matrix of at least 7 x 14 dots.
- \* Thai words are not written with spaces between them as is done in writing most Western languages. All words within a phrase or clause, or within a sentence containing a single clause, are written together without any spacing. Spacings, on the other hand, set off the end of a phrase, clause or sentence, and are therefore used in places where we normally use the comma and the full stop. Spaces are also used in circumstances such as the following:
  - to separate words in a series
  - between the first and last names of persons
  - before and after numerals



- before and after the repeat sign
  - before and after the sign which is the equivalent of "etc."
  - before and after parentheses
- \* In Thai there is a recognised lexical order which is roughly as follows:
- Each word is listed under the first written consonant, and each consonant in turn has a recognised dictionary order.
  - Consonants file before vowels, and vowels in turn have a recognised "alphabetical" order or collating sequence.
  - If identical words have different tone marks, the collating order of the tone marks is taken into consideration. Because the input data for the Thai character set includes a mix of consonants, vowels, tone marks and special symbols without the use of superscripts and subscripts, it may not be possible to sort the input data rigidly according to the Thai lexicographical order. On the other hand, sorting at the output stage requires relatively complicated processing.

The solution to the above problems which were used during the SEAPRINT Pilot Project were as follows. At the input stage, all Thai characters are input in one line of text. However, at the output stage a complicated algorithm for printing specifies that, for example, if a vowel is to be printed above a consonant and a tone mark is to be written above the vowel, the characters should be input in the order of consonant, vowel and tone mark. A printing program then provides that every line of output takes up four levels in which, for example, tone marks are printed on level one, superscript vowels on level two, consonants on level three, and subscript vowels on level four. (See Fig. 5, Example 1).

In the case of word divisions, the solution was to use the <blank> character and the special character "@". Example 2 in Fig. 5 illustrates the use of the <blank> and "@" characters.

In the example, the <blank> indicates where truncation is permitted between clauses, phrases or words. That is to say, if two words can fit into one line of text, they will be output without a blank space, i.e. the words will be joined. On the other hand, the blank space will be taken into consideration when word division is required for line editing purpose, the computer being programmed to allow word divisions only where the <blank> is indicated in the input data. A useful analogy is to liken the <blank> to the soft hyphen of a word processing program which

will be activated when an English word has to be truncated at the end of a printed line.

The "@" character may be likened to a real blank, that is to say, whenever two words are output, there must always be a blank space between them, with the "@" character suppressed at the printing stage. Each "@" character in the input stage is equivalent to a blank at the output stage.

*Mixed script processing.* Processing Thai and roman character sets simultaneously does not constitute a major difficulty, since each of the character sets have been given unique codes in the EBCDIC coding scheme devised by IBM Thailand. The only major difficulty is that the Thai characters occupy double the number of line spaces required by the roman characters. As a result of this, when there is mixed character processing, double line spacings are required for each line of text containing both Thai and roman characters. (See Fig. 6)

#### AVAILABILITY OF SOFTWARE

Three types of software are currently available or are in the process of being developed which can process Thai and roman scripts, or Jawi and roman scripts simultaneously. The first kind, which is basically for word processing, has been developed by Wang Computers. The Jawi word processor that Wang has recently announced in Malaysia is a modification of its Arabic word processing system used extensively in the Middle East. A Thai word processing system is also available. Unfortunately, however, I do not possess sufficient details to describe the word processing systems concerned.

The second type of software suitable for library use is a database management system known as MINISIS, which was developed by the International Development Research Centre in Canada. This software package has the capability of processing a wide variety of scripts, using the principle of a character switch indicator. As suggested above, the non-roman scripts are stored in separate fields. MINISIS is capable of handling Chinese and Arabic; consequently, I do not think that there would be any problems in devising input/output devices capable of inputting and outputting Thai and Jawi.

The third type, which was used for the SEAPRINT Project was based on the MALMARC software (used in Malaysia for centralised cataloguing and processing of MARC-based records). The software was used with modifications to process both Thai and roman script records. This was possible because both the scripts were given unique codes within the same coding scheme. For the same reason, it was possible to mix character sets in the same field. Certain modifications were of course required. In terms of hardware, special input and output devices were required, and in terms of software a special print program had to be written to cope with the peculiarities of the Thai character set.

## CONCLUDING REMARKS

It can be seen from the above discussion that while Jawi and Thai have certain special features which complicate their computer processing, the solutions to the problems are not difficult to find, especially as both languages are "mappable" using any of the standard coding schemes. Unfortunately, various problems still exist, and unless the will exists to solve them, they are likely to remain for a long time to come. What are these problems?

Firstly, no internationally recognised standards for coding Thai and Jawi exist. At the present time, each manufacturer that develops a system which can process a non-roman script tends to decide arbitrarily on the codes that should be used. For example, the IBM developed system for Thai appears to have been done rather hurriedly, and without much thought.

Secondly, it difficult to persuade computer manufacturers to develop hardware and software systems for which there is a limited demand. For example, Jawi is not widely used in Malaysia today, and hardly at all in Indonesia. Consequently, there would be little incentive to develop systems that can process Jawi records. It was possible to persuade IBM Thailand to produce some limited input and output devices for the Thai character set because Thai is used extensively in Thailand.

Thirdly, no software exists yet that can store, process retrieve and output Thai and Jawi records. MINISIS has that in-built capability, and it is my understanding that IDRC is working towards providing MINISIS with the capability of processing a wide variety of scripts. Most other library packages, unfortunately, were designed to recognise only roman scripts.

Because computer technology is advancing at a rapid pace, I am fairly optimistic that before too long, software and hardware will be available which will allow the computer processing of a wide variety of scripts.

lht/seascrpt.ws2

#### REFERENCES

1. Eadie, Donald (1982). *A user's guide to computer peripherals*. Englewood Cliffs, N.J.: Prentice Hall.
2. Lewis, M.B. (1954). *A handbook of Malay script*. London: Macmillan.
3. Lim, Huck Tee (1983). "SEAPRINT: a computerized regional bibliography project." In: *The library in the information revolution: proceedings of CONSAL VI*. Singapore: Maruzen Asia, 1983 pp.244-258.
4. Haas, Mary R. (1956). *The Thai system of writing*. Washington, D.C.: American Council of Learned Societies.
5. Perez, A.Q. et al. (1978). *Papers from the conference on the standardisation of Asian languages*, Manila, Philippines, Dec. 16-21, 1974. Canberra: Dept. of Linguistics, Australian National University, 1978.
6. Wilkinson, Barry & Horrocks, David. (1980). *Computer peripherals*. London: Hodder and Stoughton.

LHT/nsnl/seascrpt.ref

Fig. 1. Roman character  
set using 7x7 dot matrix





Fig. 2. Examples of Jawi character set using  
7x13 dot matrix



Fig. 3. Some Jawi alphabets

THE ALPHABET

9

THE ALPHABET

Isolated

Initial Medial Final

1.  *alif*

Note that when *alif* is isolated or initial it is a downward stroke, when it follows another letter it is an upward stroke.

*Alif* is never joined to the letter which follows it.

2.  *ba*

Note that when a 'saucer' letter is cut short the dot moves over to the right.

In the medial position the first arrow-stroke shows the preceding letter making a junction with *ba*. The dot is usually placed immediately under the notch so created.

3.  *ta*

In continuous writing the two dots would appear over the notch, in the initial and medial forms.

Fig. 3. Some Jawi alphabets (cont'd)

10

A HANDBOOK OF MALAY SCRIPT

Isolated

Initial Medial Final

4.



*tha*

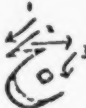


Note that (for initial and medial positions) one notch with three dots above it is *tha*, two notches with three dots above is *shin* (No. 14)—both of them uncommon—and one notch with three dots below is *nya* (No. 35).

5.



*jim*



Notice the method of joining a flat-topped letter. In the medial and final positions the first arrow indicates a joining line<sup>1</sup> coming in from the preceding letter. This line may be made to carry dots, and so count as a letter itself. This joining-line, when not used to carry a dot or dots belonging to the preceding letter, is the *only* meaningless stroke in *Jawi* script. All other strokes or curves must be accounted for as parts of letters.

6.



*cha*



In rapid writing this three-dotted arrow is easy to confuse with initial and medial *nya* (No. 35). Look for the joining line.

<sup>1</sup> In print the joining-line comes up from below, thus ج.

| Thai characters | Hexadecimal | Thai characters     | Hexadecimal |
|-----------------|-------------|---------------------|-------------|
| CONSONANTS :    | 80          | ก                   | AA          |
| ข               | 81          | ค                   | AB          |
| ฃ               | 82          | ด                   | AC          |
| ฅ               | 83          | ต                   | AD          |
| ฉ               | 84          | ซ                   | AE          |
| ช               | 85          | ฌ                   | AF          |
| ฌ               | 86          | จ                   | BO          |
| ฎ               | 87          | ฉ                   | B1          |
| ฏ               | 88          | ช                   | B2          |
| ฐ               | 89          | ฌ                   | B3          |
| ฑ               | 8A          | ฌ                   | B4          |
| ฒ               | 8B          | ฌ                   | B5          |
| ณ               | 8C          | ,                   | B6          |
| ด               | 8D          | ๖                   | B7          |
| ด               | 8E          | ๖                   | B8          |
| จ               | 8F          | ๖                   | B9          |
| ก               | 90          | ๖                   | BA          |
| ข               | 91          | ๖                   | BB          |
| ฃ               | 92          | ๖                   | BC          |
| ฅ               | 93          | ๖                   | BD          |
| ฉ               | 94          | Tonation marks :    | BE          |
| ช               | 95          | ๖                   | BF          |
| ฌ               | 96          | ๖                   | CA          |
| ฎ               | 97          | ๖                   | CB          |
| ฏ               | 98          | Special character : | CC          |
| ฐ               | 99          | ๖                   | CD          |
| ฑ               | 9A          | ๖                   | CE          |
| ฒ               | 9B          | ๖                   | CF          |
| ณ               | 9C          | ๖                   | D7          |
| ด               | 9D          | (above consonant)   | DA          |
| ด               | 9E          | (under consonant)   | DB          |
| จ               | 9F          |                     |             |
| ก               | A0          |                     |             |
| ข               | A1          |                     |             |
| ฃ               | A2          |                     |             |
| ฅ               | A3          |                     |             |
| ฉ               | A4          |                     |             |
| ช               | A5          |                     |             |
| ฌ               | A6          |                     |             |
| ฎ               | A7          |                     |             |
| ฏ               | A8          |                     |             |
| ฐ               | A9          |                     |             |

Fig. 4. Thai character set with equivalent EBCDIC codes

Fig. 5. Examples of solutions  
to the problems of  
outputting Thai script

Example 1

Data in a computer storage:

ห อ ส ม . ค ท - ป - น - ง  
↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓  
3 3 3 3 4 3 3 2 1 3 2 3 2 3

printing position

Printed in this form

1 st line  
2 nd line  
3 rd line  
4 th line

\_\_\_\_\_  
\_\_\_\_\_  
ห อ ส ม ค ท ป น ง  
\_\_\_\_\_

Example 2 With blank and @ in data

Data in a computer storage;

แ น บ ก จ - ค ท ม . @ ห อ ส ม . ค

Printed in this form

1  
2  
3  
4

↓ ↓  
แ น บ ก จ ค ท ม ห อ ส ม ค  
↓ ↓  
no blank 1 blank



Fig. 6. Sample page from SEAPRINT  
bibliography

629 ENGINEERING - OTHER BRANCHES

629.225

วิศวกรรม การชลประทาน

โครงการชลประทาน = IRRIGATION - การชลประทาน, วิศวกรรม

ความรู้ทางชลประทานฉบับที่ 4, 2524.-172 หน้า,

พิมพ์ครั้งที่ 27 ก.ค. 55

629.225

JAAFAR BIN MOHC. TAHMIM

PANDUAN KESELAMATAN PENJAJAN

JENTERA BERAT / JAAFAR BIN AHMAD

TAHRIM. - KUALA LUMPUR : JAJATAN

PERMUTANAN, 1981. - V. 02MS. 9

ILL. 30CM.

JUDUL KULIT DEPAN

PERCUMA (KKT.)

(14321-127)

629.8

LOW COST AUTOMATION, (WORKING

PAPERS). - PETALING JAYA : PUSAT

DAYA PENGELJAHAN NEGARA, 1981. -

IV. (VARIOUS PAGES), ILL.

30CM.

FREE TC PARTICIPANTS (14021-222)

630 AGRICULTURE & RELATED TOPICS

630.717072

เกษตรกรรม การชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน

เกษตรกรรมและการชลประทาน = STUDY ON

YUMAKSETAKJAN 424JERSI KHAJALEJJE

NEED IN CHANGUAT SURATTHANI.-

กรุงเทพฯ, 2524.-16, 147 หน้า, 2524.-

16, 147 หน้า, 2524.-16, 147 หน้า, 2524.-

5.5.55

631 CROPS & THEIR PRODUCTION

631.44072

การชลประทาน การชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน =

APPLICATION OF U.S. SOIL

CLASSIFICATION SYSTEM FOR SELECTED

SOILS IN ORCHARD GRADING AREA,

CHANGWAT CHANTABURI.-กรุงเทพฯ, 2524.-

13, 133 หน้า, 2524.-13, 133 หน้า, 2524.-

13, 133 หน้า, 2524.-13, 133 หน้า, 2524.-

5.5.55

631.46072

ดิน การชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน

การชลประทานและการจัดการน้ำในเขตชลประทาน =

GEOPHYSICAL LAND CLASSIFICATION

AND ASSESSMENT FOR MANAGEMENT

PLANNING AT PHUPHAN NATIONAL PARK,

SAKUNNAKHON PROVINCE.-กรุงเทพฯ, 2524.-

13, 125 หน้า, 2524.-13, 125 หน้า, 2524.-

13, 125 หน้า, 2524.-13, 125 หน้า, 2524.-

5.5.55



1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

64-MCULTP-4-E  
INF

# Use of Arabic Script in Computerized Information Systems

By

Dr. Mohammed M. Aman  
Dean and Professor  
School of Library and Information Science  
University of Wisconsin-Milwaukee  
Milwaukee, Wisconsin  
U.S.A.

KEYWORD(S):

For internal use only:

To be brought to meeting no:

Friday 22 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

アラビア文字: アラビア文字は、約50ヶ国10億近くの人々に使われている。右から左へと書かれるアラビア文字は、イラン、北インド、パキスタン(一部)の言語であるFarsiでも使用される。アラビア語とベルシャ語の書法を修正したものがウルドゥ文字である。Farsiと同様、ウルドゥ語のアルファベットの37字(アラビア語より9字多い)と多くの上付き、下付き文字を持っている。本文第2ページの文字は正規のアラビア文字にはないが、母音は同じである。音声符号が多様なのは、ウルドゥ語に借入されたアラビア語とベルシャ語がその綴りの原型を保持していることに由来している。

アラビア語の文字は、ほとんど筆記体の字で連結され、いくつかの書法が今日使用されている。最も一般的なものが、Naskh'とNaskh'liqである。アルファベットに加え最低5つの母音がある。書法における母音の使用は任意であるが、文脈上から正確な語を類推できない場合には使用される。情報科学、検索、人工知能、言語理論の分野では、母音を無視することはできない。アラビア文字、その名称と英語の発音(表1)、アラビア、ベルシャ、ウルドゥ語使用される母音の位置ないし発音記号(表1a)を付す。

#### 自動化された情報処理におけるアラビア文字の使用 : 3つの問題

1. 文字表現: 1976年以来、アラブ世界のコンピュータ専門家は、情報科学におけるアラビア文字の統一コードを作るためにさまざまな分野の専門家と共同作業をしてきた。その結果、ASMOのTechnical Committee No.8が準備した統一コードNo.449-1982が制定されている。コード表現はISO-646に示唆を受けているが、ほかに制御文字やグラフィック文字セットを規定している。グラフィック文字セットによってテキストが母音はどう取り込んでいるかの判定が可能になる。

ASMOの基準は、特殊なアプリケーションでの情報交換や文字セットが不十分な場合に拡張性をもたらすものとなる。基本コード表(表1)は、ISO-646で示されているラテンの文字セットを分離し、コンピュータを使う際に完全にアラビア文字を処理できるようにアラビア語での基本プログラムを使用する形態に修正された。この基準を適用するためには、MTやネットワーク上での転送をする際のいくつかのアプリケーションを必要とする。

アラビア文字基準の範囲: アラビア文字基準は、128の文字セットからなる。そのほとんどは変更不可であるが、国別の事情もあり、柔軟性を持たせる規程がつけられている。

2. キーボードの問題: ローマ字とアラビア文字双方を扱うターミナルが不可欠である。アラビア文字用に英語の小文字(lower case)を使えなくしている。
3. アラビア語の連結文字の問題: ワードプロセッサを改良をして、キーボード上に文字をすべて配列する必要をなくした。コンピュータ・メーカーの大手はほぼ、今日までアラビア語処理に参入していない。いくつかのメーカーでは、アラビア文字データを取り込めるようにOSを改良しようとしている。Xeroxは、文字セットを拡張するプロジェクトに乗り出している。

アラビア文字コレクションへのコンピュータ利用: 主にアラビア文字資料を所蔵している図書館での自動化計画の場合も、他の図書館と同様なアプリケーションである。が、アラビア諸国の図書館員や自動化の専門家は、すでに制定されている基準に従っていない。例えば、MARCフォーマットはないし、ASMO基準は多くのシステムに適用されていない。さらにアラビア人の名称の検索を容易にするような計画もない。

市販ソフトのアラビア語化: クウェートでは、NSTICがStairsのアラビア語バージョンを使用している。

Stairsが、アラビア語のコマンドとブール演算によって検索できるように開発されつつある。カナダのIDRCが開発したパッケージであるMINISISのアラビア・バージョンが、1982年にALDOCによって完成された。ほかにもVTLSやDOBIS/LIBLISのようなソフトウェアのアラビア語化の研究が進行中である。

Use of Arabic Script in  
Computerized Information Systems

Abstract of A Paper  
Presented by

Dr. Mohammed Aman  
Dean and Professor  
School of Library and Information Science  
University of Wisconsin-Milwaukee  
Milwaukee, Wisconsin  
U. S. A.

At the IFLA '86 Preconference Seminar on  
Automated Systems for Access to Multilingual and  
Multiscript Library Materials

Application of computer technology in libraries whose predominate collections are in Arabic Script is a new development which will have a wide impact on the operations and procedures of these libraries.

The paper explains the complexity of Arabic Script, and the levels of solutions to the problems of using Arabic characters in computerized information processing. The well known Arabic Standard No. 449-1982 is explained and its relation to ISO 646 is pointed out.

Automated programs in libraries whose holdings are mainly in Arabic script parallel similar applications in other libraries and information centers.

The paper describes the experiences of Arab libraries in automating their procedures and services and in Arabizing off-the-shelf software.

### The Arabic Script:

The Arabic script is used by nearly one billion people in approximately fifty countries. The Arabic characters, which are written from right to left, are used in the Farsi, the language used in Iran, northern India and also in some parts of Pakistan. A modified form of Arabo-Persian writing system is used as the Urdu script. It utilizes all the letters of the Persian script along with few additional symbols for sound peculiar to Urdu.

Like the Farsi, the Urdu alphabet contains 37 letters (about nine letters more than in the Arabic language) and a number of supplementary superscripts and subscripts. All Arab consonants have separate symbols.

The following letters are not represented in the regular Arabic alphabet, but the vowels are the same:

پ pee (p) Farsi pe

ت tee (t)

چ cee (c) Farsi che (ch)

د daal (d)

ز zee (r) Farsi zhe (zh)

گ gaaf (g) Farsi gaf (g)

ع barii yee (ee)

The multiplicity of symbols for certain sounds arises from the fact that Arabic and Persian words borrowed into Urdu preserved their original spellings although their pronunciation was adaptive to the native sound system. ت "tee," س "siin," and چ "chootii hee" occur in both Indo-Aryan and Perso-Arabic words. Words borrowed from languages other than Arabic and Persian always have "tee," "siin," "chootii hee," for t, s, h and "zee" for z.

Letters in the Arabic word are joined together in a highly calligraphic script. Several styles of calligraphy are in use today. The



most common of these are called "Naskh" and Nasta'liq." When Arabic printing was introduced early in the eighteenth century, the shapes of Arabic letters were an imitation of handwriting which was calligraphic. Thus, at first, printing was an imitation of calligraphy, even if little simplified. As literacy spread, the opposite became true; writing became a simplification of printing!(1) This process continued and still continues.

The same alphabet in Arabic has different forms depending on that letter's relative position in the word. If the letter comes in the beginning of the word, it is having one form **ا**, when it is in the middle of the word, it is another form **آ**, and if it is the last letter of the word, it is entirely different **آ**, as in **آآآ** whereas in English there are only two forms-capital letters (also known as upper case letters) and small letters (lower case letters). The horizontal distance taken by one letter varies considerably from one to the other. Sometimes one character is double the size of the other.

Besides the alphabet character set there are at least five vowels. These are optional in writing, used only when it is not possible to guess the exact word from the context. They are used in conjunction with other letters printed above or below the letter. In complex applications in the field of information science, search and retrieval, artificial intelligence, and language theory it is impossible to ignore them.(2)

Table 1 lists Arabic letters, their names and pronunciation in English, and Table 1a shows vowel points or diacritical marks used in Arabic, Persian and Urdu.

**TABLE 1**  
**Arabic Letters, Their Names and Pronunciation**

|                                      |                        |
|--------------------------------------|------------------------|
| 1 ء hamza: glottal stop              | 17 ﺩ dad: emphatic d   |
| 2 ا alif: a                          | 18 ﺕ ta': emphatic t   |
| 3 ب ba': b                           | 19 ﺫ dha': emphatic dh |
| 4 ﺕ ta': t                           | 20 ﺀ 'ayn              |
| 5 ﺕ ta'maqsurah: t or h              | 21 ﺥ ghayn: gh French  |
|                                      | Pronunciation of r     |
| 6 ﺕ tha': th as in thin              | 22 ﺕ fa': f            |
| 7 ج jim: j                           | 23 ق qaf: q uvular k   |
| 8 ح ha': Spanish pronunciation ha    | 24 ك kaf: k            |
| 9 خ kha': German pronunciation of ch | 25 ل lam: l            |
| 10 د dal: d                          | 26 م mim: m            |
| 11 ذ thal: dh as in dhu'l-Hijja      | 27 ن num: n            |
| 12 ر ra': r                          | 28 ه ha': h            |
| 13 ز za': z                          | 29 و waw: W            |
| 14 س sin: s                          | 30 ي ya': y as in yes  |
| 15 ش shin: sh                        | 31 ﺀ alif-maqsurah: a  |
| 16 ﺀ sad: emphatic s                 |                        |

**TABLE 1a**  
**Vowel Points**

|      |                     |
|------|---------------------|
| 32 َ | fathah: (a)         |
| 33 ُ | dammah: (u)         |
| 34 ِ | kasrah: (i)         |
| 35 ّ | shaddah: (doubling) |
| 36 ْ | sukun: (stop)       |

### Use of Arabic Script in Computerized Information Processing:

There are three levels of solutions to the problems of using Arabic characters in computerized information processing:

#### 1. Character Representation

Since 1976 computer experts from the Arab World worked together and with other information, computer and language experts to develop a unified code for the use of Arabic script in informatics.(3) The result was the unified

Arabic standard specification no. 449-1982 which was prepared by Technical Committee no. 8 (Arab Characters in Informatics) of the Arab Standards and Metrology Organization (ASMO). Among the parties who participated in its preparation are the Arab League ,Educational, Cultural, and Scientific Organization (ALECSO), and the Institute of Studies and Research for Arabization in Morocco.

The Arabic standard specifies the properties of a coded character set using 7 bit binary codes for information interchange among different types of data processing equipment using the Arabic characters. It also specifies a set of control and graphic characters in addition to its coded representation inspired from ISO 646. The set of specified graphic characters in this standard makes it possible to represent Arabic text whether it is totally vowelized, partially vowelized, or unvowelized.(5)

The ASMO Standard provides the possibilities for information interchange for special application, as well as the possibilities for expansion in case of the insufficiency of the coded character set.

The Arabic Standard was made in accordance with ISO 646, and some points were modified so that the standard ISO 646 is convenient for Arabic usage.

The basic code table (Table 1) was modified in such a way which permits the usage of the coded character set as a separate group from the Latin character set described in ISO 646 for information interchange, and the use of basic programs in Arabic language for the purpose of complete Arabization when using computers. The table also allows the usage of the coded character set together with the Latin character set because of the correspondence between these two standards.

Table 1

Basic Code Table  
ASMO Standard Specification 449

|    |    |    |    |     |                          |               |                |                |   |                |   |                |                |  |
|----|----|----|----|-----|--------------------------|---------------|----------------|----------------|---|----------------|---|----------------|----------------|--|
|    |    |    |    |     | b.                       | 0             | 0              | 0              | 0 | 1              | 1 | 1              | 1              |  |
|    |    |    |    |     | b.                       | 0             | 0              | 1              | 1 | 0              | 0 | 1              | 1              |  |
|    |    |    |    |     | b.                       | 0             | 1              | 0              | 1 | 0              | 1 | 0              | 1              |  |
|    |    |    |    |     | 0                        | 1             | 2              | 3              | 4 | 5              | 6 | 7              |                |  |
| b. | b. | b. | b. | row |                          |               |                |                |   |                |   |                |                |  |
| 0  | 0  | 0  | 0  | 0   | NUL                      | TC,<br>(BLE)  | SP             | 0              | ا | ذ              | - | -              | -              |  |
| 0  | 0  | 0  | 1  | 1   | TC,<br>(SDM)             | DC.           | !              | 1              | . | ر              | و | -              | -              |  |
| 0  | 0  | 1  | 0  | 2   | TC,<br>(STX)             | DC.           | "              | 2              | ت | ز              | ق | .              | .              |  |
| 0  | 0  | 1  | 1  | 3   | TC,<br>(ETX)             | DC.           | # <sup>②</sup> | 3              | ا | س              | ص | ③              | ③              |  |
| 0  | 1  | 0  | 0  | 4   | TC,<br>(EDT)             | DC.           | □ <sup>②</sup> | 4              | ف | ش              | ر | ⑤              | ⑤              |  |
| 0  | 1  | 0  | 1  | 5   | TC,<br>(END)             | TC,<br>(INAK) | %              | 5              | ! | ط              | م | ⑤              | ⑤              |  |
| 0  | 1  | 1  | 0  | 6   | TC,<br>(ACK)             | TC,<br>(SYN)  | &              | 6              | ذ | ض              | ز | ⑤              | ⑤              |  |
| 0  | 1  | 1  | 1  | 7   | BEL                      | TC,<br>(ETB)  | '              | 7              | ا | ط              | ه | ⑤              | ⑤              |  |
| 1  | 0  | 0  | 0  | 8   | FE,<br>(BS)              | CAN           | ) <sup>④</sup> | 8              | . | ظ              | و | ⑤              | ⑤              |  |
| 1  | 0  | 0  | 1  | 9   | FE,<br>(HT)              | EM            | ( <sup>④</sup> | 9              | ة | ع              | ى | ⑤              | ⑤              |  |
| 1  | 0  | 1  | 0  | 10  | FE, <sup>①</sup><br>(LF) | SUB           | *              | :              | ز | غ              | ي | ⑤              | ⑤              |  |
| 1  | 0  | 1  | 1  | 11  | FE, <sup>①</sup><br>(VT) | ESC           | +              | : <sup>④</sup> | ذ | ]              | - | ④              | ④              |  |
| 1  | 1  | 0  | 0  | 12  | FE, <sup>①</sup><br>(FF) | IS,<br>(FS)   | .              | > <sup>④</sup> | ج | \              | " |                |                |  |
| 1  | 1  | 0  | 1  | 13  | FE, <sup>①</sup><br>(CR) | IS,<br>(GS)   | -              | =              | > | [ <sup>④</sup> | - | { <sup>④</sup> | { <sup>④</sup> |  |
| 1  | 1  | 1  | 0  | 14  | SO                       | IS,<br>(RS)   | .              | < <sup>④</sup> | خ | ^              | - | -              | -              |  |
| 1  | 1  | 1  | 1  | 15  | SI                       | IS,<br>(US)   | /              | ? <sup>④</sup> | د | -              | - | DEL            | DEL            |  |

Applying this standard requires several application standards to be implemented on a carrier (magnetic carrier, transmission network, etc.). These applications are specified in other ASMO standards.

### **Scope of the Arabic Standard:**

The Arabic Standard contains a set of 128 characters (control characters and graphic characters such as letters, digits and symbols) with their coded representation. Most of these characters are mandatory and unchangeable, but provision is made for some flexibility to accommodate special national and other requirements.(5)

The Arabic character set is primarily intended for the interchange of information among data processing systems and associated equipment, and within message transmission systems.

The advantage of the set stems from its being applicable to all Arab alphabets, and its flexibility for extension where its 128 characters are insufficient for particular applications.

### **2. Problems of the Keyboard**

The second level of solutions has to do with the problems of the keyboard of the computer. Terminals that can handle both Roman and Arabic scripts are now a necessity in computer configurations or in countries which use Arabic and Arabo-Persian characters. To modify the existing keyboard many extra keys have to be provided. The result has been that small or lower case English letters have disappeared from the keyboard. Arabic letters occupy the same internal notations as of the lower case English characters.

### **3. Problems of Connecting Characters in the Arabic Word ,**



To solve the third problem of connecting characters in the Arabic word, some word processors have adopted the techniques of Contextual Analysis. By adopting this technique it is not necessary to keep all the various forms of one letter on the keyboard. The system will automatically pickup the correct form depending on its relative position within the word.

The problem of variances in the horizontal distances taken by each letter has been solved by many word processors who have made their screen as a proportional screen instead of fixed pitch. But this feature has not been successfully tackled in any computer so far.

Most of the large computer manufacturers have not stepped into the Arabization Process so far. But those who have done something about it, have changed the operating system to incorporate entry of Arabic data. In most cases the machine works in two modes.

For the Arabic mode, the manufacturers realize that while the Arabic alphabet runs from right to left, numerals run from left to right. Some manufacturers have entered Arabic numbers from right to left, and internally being converted otherwise.

Some companies like Xerox are developing projects to expand their character code Standard for the Arabic alphabet using international Standards and ISO proposals. In the coding process, Xerox has also included Indian numerals and Farsi characters.

### **Use of Computers in Processing Arabic-Script Collections:**

Automated programs in libraries whose holdings are mainly in Arabic script could parallel similar applications in other libraries and information centers. These applications could be grouped as follows:

- 1.<sup>2</sup> Bibliographic control systems

2. Circulation control
3. Indexing systems
4. Word processing
5. Database development

In spite of attempts made by librarians and automation specialists in countries where Arabic script is used, many of them do not follow established standards. For example, there is no MARC format for Arabic materials, and the ASMO standards are not applied in many systems. Furthermore, there is no standard scheme for representing Arabic names to facilitate subsequent searching in a DBMS environment: First, Middle, Last; Last, First, Middle; Initials, Last, etc.

The Arabic language is endowed with a grammar that is universally considered to be intricate and complex. Despite this, it affords the linguist a comprehensive and extremely well-defined set of rules for the derivation of complex words from their simple roots. The exploitation of these rules in the Arabic information processing environment has yet to be seriously investigated.

#### **Arabizing Off-the-Shelf Software:**

Arab librarians have made considerable efforts to Arabize off-the-shelf software to accommodate their needs and requirements for the storage and retrieval of Arabic documents.

In Kuwait, for example, the National Scientific and Technical Information Center (NSTIC), has introduced an Arabic version of STAIRS. STAIRS/CMS allows the user to modify the system to accept non-English characters. By Arabizing the keyboard, commands, and operands and increasing

the character set to include the Arabic characters, a version of STAIRS has been developed that allows a user to retrieve Arabic documents using Arabic commands and Boolean operators.(6)

MINISIS, a software package developed by the Canadian International Development Center (IDRC) for Hewlett-Packard (HP) 3000 minicomputer, has been Arabized by the Arab League Documentation Center (ALDOC) in 1982. The Arabic version has been used in other Arab libraries and information centers such as the Arab League Educational, Cultural and Scientific Organization (ALECSO) in Tunis, Tunisia, National Documentation Center (CND), Rabat, Morocco, National Information Center for Finance and Industry (NICFI), Riyadh, Saudi Arabia, to mention a few. Research is being conducted on the arabization of other software such as VTLS (Virginia Tech Library System) and DOBIS/LIBIS system.

Results of research in the handling of Arabic script will also benefit others interested in developing truncation rules of Arabic words for internal storage and retrieval, and developing sophisticated Arabic data compression algorithms for both storage and communication (transmission) of Arabic information.

### References

1. Aman, Mohammed M. Use of Arabic in Computerized Information Interchange. Presented at the IFLA 49th Conference, Munich, W. Germany, 1983. p.4.
2. Akhtar, Sayeed. "Arabization: no one has tackled it seriously yet," Saudi Computerworld, 1(3): March 1984,

3. Aman, Mohammed M. "Use of Arabic in Computerized Information Interchange," Journal of the American Society for Information Science, 35(4):July 1984, p.205.
4. Arab Standard Specifications No. 449-1982: Data processing-7 Bit Coded Arabic Character Set for Information Interchange. Amman, Jordan: Arab Organization for Standardization and Metrology (ASMO),1982.
5. Ibid. p. 1.
6. Khalid, F. "Automation in a Special Library in Kuwait," Information Technology and Libraries, 2(4):Dec. 1983, p.360.



1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

65-MCULTP-5-E  
INF

Development of Greek Automated Cataloguing System

Victoria, Australia

By

Eve STOCKER

Coordinator Non English Language Cataloguing Unit

Technilib

Melbourne, Australia

KEYWORD(S):

Automated systems, Greek  
Greek catalogues  
Greek cataloguing  
Non-Roman scripts

For internal use only:

To be brought to meeting no:

Friday 22 August 1986

Remarks:

to be presented orally ☒

for distribution only ☐



1977年から1978年にかけて、ヴィクトリア州メルボルンの2つの公共図書館—キャリングブッシュとハイデルベルク地域図書館—は、ギリシャ語で書かれた図書をギリシャ語で目録する自動化システムの構築を検討するために会合をもった。ギリシャ語資料の自動化目録作成システムは、英語およびその他のローマ字を使用する言語資料に対する目録・貸出業務用等の図書館自動化システムの論理的展開とみなされていた。1975年以来、これら2つの公共図書館は、ギリシャ語その他のコミュニティ言語による蔵書を構築してきた。([「コミュニティ言語」という語句は、非英語使用国から来た移民の母国語を表現するのに用いられる。])

この一世界初のギリシャ語資料目録作成システムの開発は、Libramatic System社という、ヴィクトリア州にある各種の図書館、機関、図書館サービス会社を顧客とするデータ処理サービス会社によって引き受けられた。この開発が記念碑的なプログラム作成の仕事であることが後に明らかになったのである。

まず最初に、大文字・小文字双方のギリシャ・アルファベットの文字セットを考案し、その個々の文字の定義を、特別に改良したBeehive B550端末にプログラムし組み込む必要があった。ギリシャ文字はこの端末の標準ローマ字キーボード上に刻印された。画面上には英語のアルファベット、すべてのALA文字セット、そしてスーパーシフトキーの打鍵によってギリシャ文字が表示される。

ギリシャ語レコードには標準MARCフォーマットを使用したのであるが、ギリシャ語のデータとそれに対応する英語のデータ(LC件名標目)もしくは翻字形のデータ(著者、標題)という形で並列フィールドを保有している。ギリシャ語レコードの翻字形の部分は貸出システム用に必要だったのである。Libramatic System社は、ギリシャ語資料のマイクロフィッシュ目録を3つの配列のもとに作成した。すなわち著者/標題、英語の件名標目(LCSH)、およびギリシャ語の件名標目(翻訳LCSH)である。また、書架リスト、所在リスト、校正作業リスト(Validation Report)(以上マイクロフィッシュ)と差し込み票、背ラベルが作成された。最初のマイクロフィッシュ目録は1981年に作成されたのである。マイクロフィッシュ作成用にギリシャ語文字セットを定義するにあたって、端末をギリシャ文字が表示できるように改良したり、ギリシャ語レコードが処理可能のようにプログラムを修正した際に経験したのと同様の問題に突き当たった。

1983年までには、ギリシャ語資料目録作成ネットワークは5つの公共図書館で構成されるまでになり、目録作業の量も大幅に増大していった。

1984年にギリシャ語資料目録作成ネットワークは、その目録業務をTechnilibの目録サービス部門に吸収するよう提案した。Technilibは、1975年にヴィクトリア州図書館評議会からの助成金を得て、公共図書館グループにより共同で組織された集中目録作成機関である。Technilibは、それまで(非ローマ字に対しては翻字を用いて)多数の言語の図書を目録してきており、ギリシャ語資料の目録業務もその他の非英語資料目録作成サービスとの間に整合性のあるものとみなされた。Technilibの目録作成およびその他のデータ処理サービスは、当時Libramatic Systems社によって提供されていたのである。

1985年にTechnilibは、IBM4381汎用機の上でDobis-Libsを用いた部内のオンライン目録業務システムを導入し、Libramatic Systems社のバッチ処理システムの利用を段階的に取り止めていった。その前年にLibramatic Systems社は、Idapsという別のデータ処理会社と合併し、Idaps Australia社のLibramatic部門となっていた。TechnilibがIdaps社から撤退して以後、ギリシャ語資料の目録業務に関わる処理作業の料金は値上げされていった。1986年にTechnilibでは、ギリシャ語資料の目録業務を現行のシステムに吸収するという方針が決定された。それにはギリシャ文字を表示できるようにEricsson Alfaskop端末を改良し、プログラムの修正を行うことが必要になる。

ABSTRACT: DEVELOPMENT OF GREEK AUTOMATED CATALOGUING SYSTEM  
VICTORIA, AUSTRALIA

Starting in 1977-78 what is believed to be the first automated cataloguing system for Greek books, in Greek, was developed by two public libraries in conjunction with a library data processing service in Melbourne, Australia. This involved designing a Greek character set, modifying a data entry terminal to display Greek and extensively adapting an existing automated system to process Greek records (based on standard MARC format). Similar character definitions and programming solutions were required to produce microfiche catalogues, shelf and location lists and labels. This paper also discusses administrative aspects of the Greek Cataloguing Network.

Eve STOCKER  
Coordinator Non English Language Cataloguing Unit  
Technilib  
Melbourne, Australia

## DEVELOPMENT OF GREEK AUTOMATED CATALOGUING SYSTEM

### VICTORIA, AUSTRALIA

#### Brief background: immigration to Australia

Up to the time of the second world war the majority of immigrants to Australia came from the United Kingdom. After this time immigration from countries in which English is not the mother tongue (particularly Northern, Eastern and Southern Europe, including Greece) increased significantly. Since the 1970s large proportions of Australia's immigrants have come from countries such as Turkey, Lebanon and other Middle Eastern countries, Chile and other South American countries, Vietnam and other Asian countries in addition to European countries.

Second and third generation Australians of many national and cultural backgrounds, as a conscious and positive decision, have retained their first language and many aspects of the cultural, including religious, life of their countries of origin.

The Australian Government has made efforts to review services and programs to immigrants (e.g., Report of the review of post-arrival programs and services for migrants, 1978; Evaluation of post-arrival programs and services, 1982), and to develop policies (e.g., National Language Policy) to determine what special needs Australia's immigrants have, whether those needs are being adequately met, and whether people's rights are being safeguarded.

The proportion of Australia's population born overseas is estimated to be 21.8% overall in 1986. The proportion of overseas-born population varies considerably from one to another of Australia's six states and two territories. By mid-1986 it is estimated that Victoria will be the home of over 44% of the Greek-born population within Australia (in numbers, just over 69,500),

and New South Wales, of 36% (just over 56,500). Almost all of the Greek born population, and the second generation of Greek born parents, in Victoria, live in the state's capital, Melbourne. (AIMA 1982, pp. 22-35)

Since 1975 many public libraries have made efforts to meet the needs of users from non-English speaking backgrounds by building up collections in people's mother tongues (called 'community languages') guided, in part by demand from users, and in part by availability of books in community languages.

This background provides a context for the developments which have taken place within the last 10 years in Melbourne to create an automated system to catalogue, in Greek, books in the Greek language.

#### The Greek Cataloguing Network

The idea for a computerised catalogue, in Greek, arose as the logical extension of automated cataloguing systems which had been implemented for collections in the English language and other languages which used the Roman alphabet (Italian, Vietnamese, Polish, etc.). Cataloguing of books in Greek (and other languages using non Roman alphabets) had been dealt with by creating brief transliterated records. This was considered unsatisfactory for both Greek speaking library users and library staff who, for the most part, did not speak Greek, but spoke only English.

Two public libraries within the Melbourne metropolitan area, Carringbush Regional and Heidelberg Regional, had acquired large collections of books in Greek. Ideas and recommendations for developing automated cataloguing systems for library collections in languages other than English had been discussed at a number of professional meetings examining the questions of providing multicultural library services. In part as a result of the particular individuals involved and their commitment to

multicultural library services, Carringbush and Heidelberg libraries got together to seriously consider the requirements for turning the idea of a Greek cataloguing system into a reality.

The automated system used by Carringbush and Heidelberg Regional Libraries at the time of the initial moves to automate cataloguing in Greek was provided by Libramatic Systems Pty. Ltd. Libramatic Systems provided a batch cataloguing service, and record processing service, with microfiche, print or card output. Cataloguing and circulation systems for a number of public libraries throughout Victoria were provided by Libramatic Systems under the coordination of a centralised cataloguing agency called Technilib which was set up by the Library Council of Victoria as a cooperative organisation. Libramatic Systems also provided batch work for other libraries, and other organisations (e.g. Cooperative Action by Victorian Academic Libraries, CAVAL), and had been the main library data processing service for some years. Technilib had been cataloguing Greek books for public libraries using brief, transliterated records because there was no way to create records in the Greek alphabet. The development of the Greek cataloguing system under the direction of Carringbush and Heidelberg was independent of the services Technilib provided.

Development of the Greek cataloguing system proved to be a very great challenge, as, at that time, there were no existing character generator EPROMs enabling the display of the Greek alphabet, both upper and lower case. As far as Libramatic Systems could determine, their development of a computerised Greek cataloguing system constituted a world first.

The existing cataloguing system had to be substantially adapted for the Greek system. The myriad complexities and



interrelationships involved in adapting a cataloguing system to make it work in a different language made this a monumental programming task. Some of the considerations which required particular programming solutions were: correct filing order, cataloguing requirements, sort forms, generation of call numbers to meet individual library specifications, correct sorting and printing of records on microfiche, generation of Greek lettering on labels. Close liaison was required among cataloguers, systems consultants and programmers at the libraries and Libramatic Systems. Extensive trialling with test data and real bibliographic records was required to encounter as many variations as possible. Each test raised new problems and new situations not previously encountered which then had to be solved by further programming efforts.

The development of the Greek cataloguing system began in 1977-78; in late November 1981 Libramatic Systems announced the availability of the Greek cataloguing system. Although Libramatic Systems had actually already processed some data and produced the first microfiche catalogue and validation reports two weeks earlier, there were not yet sufficient records in the system to test it adequately.

#### The form of the Greek record

The Greek bibliographic record was based on MARC record format. Data in Greek was entered in the usual MARC record fields. Subject access was provided in English as well as Greek. In addition, certain local fields were specified for the entry of data in transliterated form. The transliterated fields were required for transfer to circulation systems which were unable to handle Greek characters.

The main part of the Greek records was set up in the following way, using MARC tags as indicated, for Greek, for transliterated, and for English data:

| Greek tags | Transliterated (T)<br>or English (E) tags | Meaning                                                                 |
|------------|-------------------------------------------|-------------------------------------------------------------------------|
| 100        | 960 (T)                                   | Main Entry, personal name                                               |
| 110        | 961 (T)                                   | Main Entry, corporate/geographic name                                   |
| 245        | 962 (T)                                   | Title and statement of responsibility (962 = transliterated title only) |
| 650        | 963 (E)                                   | Topical subject heading                                                 |
| 651        | 964 (E)                                   | Geographic subject heading                                              |

Other elements in the record were: edition, publisher, physical description, series title (all entered in Greek); language and country of publication codes (using the standard AUSMARC codes) and BDC number. (See copy of update report, Attachment 1.)

#### How the system worked

As mentioned above, cataloguing was done through a batch system. The elements of the bibliographic record were written by hand onto a specially designed coding sheet, using Greek, English or transliteration in the appropriate fields.

Data was then entered from the worksheets to a specially modified Beehive B550 terminal for data collection and storage. The Greek character set was accessed through the use of a standard keyboard with a supershift key. The keyboard itself had the Greek alphabet engraved on the standard Roman alphabet keys. In addition to the Greek alphabet the terminal could display the entire ALA character set. The Greek characters which displayed on the terminal were designed by analysing a pleasant looking Greek alphabet typeface and drawing each letter onto a grid of small

squares representing the screen's display area for one character. This graphic representation then had to be translated into a program to define each character's shape on the screen in terms of pixels on/off. Additional programming was required to define the values of the characters, retaining their correct sequence in the alphabet. The Greek characters had values parallel to diacritic marks, except for the diacritics which were required in Greek coding (macron, umlaut, square brackets, etc.).

Data was stored on 8" floppy disks. When disks were ready for processing they were sent to Libramatic Systems to be read to magnetic tape and processed. Processing was carried out on a National Advanced Systems AS 5/3 mainframe. Translation programs had to be developed to turn the data on the disks into a form readable by the mainframe.

A microfiche catalogue was produced, containing three sequences: (1) author/title, (2) subject headings in English (LCSH) and (3) subject headings in Greek (translated LCSH). Validation reports, shelf lists, location lists and book labels (insert and spine) were also produced. (See examples of shelf list and labels, Attachment 2.) A similar problem was encountered in producing a Greek display for the microfiche as there was in defining Greek characters for the B550 terminal. Again, a Greek character set had to be designed and defined. The character generator ROMS in the microfiche machine had to be read, analysed and re-burned to produce Greek characters on the microfiche.

In June 1982 Libramatic Systems announced the availability of its Greek cataloguing system. It was well received, with articles appearing in both the Australian Greek- and English-language general press, as well as specialist library and computer publications.

### Growth of the Greek Network

Carringbush Regional Library remained the administrative headquarters of the Greek Network throughout the lengthy but exciting period of development and testing. Additional libraries joined the Network: Box Hill-Doncaster (in 1982), Sunshine-Melton and Moonee Valley Regional (in 1983) -- all libraries in the greater Melbourne metropolitan area. Specific call number generation programs had to be developed and tested for each new library joining the Network. In 1983 work continued on the development of programs to generate spine labels according to the specifications of individual libraries. Some required prefixes and other letters in the call number in Greek, others required English prefixes and the first three letters of the call number in romanised, transliterated form.

At this time Carringbush was coding and entering the data for all libraries in the Network except Heidelberg, which was coding and entering its own records. With the growth of the Network, pressure on Carringbush staff and resources increased. Notes from a Greek Network meeting in December 1983 indicated that a backlog existed at Carringbush of several thousand books.

### Further developments

Technilib was set up in 1975 by a group of public libraries with the assistance of a grant from the Library Council of Victoria. It is a cooperatively organised, central agency to acquire (but not select), catalogue and end-process books and other materials for public libraries throughout the state of Victoria.

By the end of 1983 the Greek Network proposed that Technilib take over the operation and administration of the Greek Network. Technilib agreed as it was felt that the Greek cataloguing work was congruent with the other non English cataloguing services Technilib provided. As mentioned, Technilib was at that time using an automated system

provided by Libramatic Systems. Administration of the Network was transferred to Technilib in August 1984.

In 1984 Libramatic Systems was merged into another data processing corporation, Idaps. It subsequently became Idaps Australia Limited, Libramatic Division, and moved its main office from Melbourne to Sydney.

In early 1985 Technilib began to install an in-house, on-line cataloguing system using Dobis-Libis on an IBM 4361 mainframe, and phased out use of the Idaps batch systems. After Technilib withdrew from Idaps, charges were raised, on a cost-recovery basis, for all processing work related to Greek cataloguing. This meant it became questionable, economically, for Technilib to continue having Idaps process the Greek cataloguing data. However, there were several processing jobs which remained to be done which were thus caught in 'limbo' by the costing changes. Several disks containing Greek data awaited processing, an up-dated catalogue was needed, and several hundred books awaited the generation of labels.

The impasse was resolved early in 1986. All data which had been entered onto disks since the previous catalogue (produced in December 1984) was processed: a microfiche catalogue, and a comprehensive update report (on microfiche) containing all the records in the

database, were produced. Labels were also generated, but because Idaps no longer had a Greek alphabet printer it was expected that names and titles would appear on the insert labels in transliterated form. Due to unforeseen difficulties, the labels were ultimately produced with nothing in the author and title fields.

#### Future developments

Greek Network members confirmed the decision in April 1986 that Technilib integrate the Greek cataloguing system into Dobis-Libis by



implementing the Greek alphabet into an Ericsson Alfaskop terminal which would require modification to enable the Greek and English alphabet, and the complete ALA character set, to be accessed. A similar process is taking place to that which was carried out by Libramatic Systems initially: designing a Greek character set, defining the display characters, and creating translation programs to enable the system to process the Greek data. With the benefit of past work to automate cataloguing in non Roman alphabets, particularly Libramatic Systems' development of the Greek Network -- equipment, programming and systems -- this should prove to be a much less daunting task than it was in 1977-78.

#### Details of the proposed system

It is planned that the Greek record will consist of full details in Greek, plus a parallel brief (circulation) record containing transliterated title, transliterated author's name, subject headings in English, classification number, dates, intellectual level, language code, country of publication code, fiction/non fiction code. The insert and spine labels will also be generated from the brief transliterated record. These two records, the Greek residing in a local file, the English in a system file, will be linked by a unique, assigned identifying number.

Until the Greek character set is implemented in a terminal, Technilib Greek cataloguing staff will be entering the brief records on line and the full Greek records onto a data sheet for entry in future when the Greek terminal is available.

Eve Stocker, Non English Cataloguing Unit  
Technilib, 123 Dover Street  
Melbourne, Victoria 3121, Australia  
19 May 1986

## REFERENCES

- Australian Institute of Multicultural Affairs. Evaluation of post-arrival programs and services. (Melbourne, AIMA, 1982).
- Greek Cataloguing Network files: Correspondence, Minutes of meetings, Non English Language Hellenic System, Non English Language Cataloguers, 1977-1986. Held at Technilib, Richmond, Victoria, Australia.
- Holmes, Anne, 'Supply and organization: the main problems in the provision of non-English materials', paper presented to the National Public Libraries Conference, Monash University, 4-7 November 1983.

## SELECT BIBLIOGRAPHY

- Anderson, Dorothy, 'ISBD in non-Roman scripts', Unesco journal of information science, librarianship and archives administration 3 (3) 1981, pp. 178-81.
- 'Greek catalogues', Multicultural libraries newsletter 2 (4) 1982, p. 19.
- 'Hellenic catalogues', New South Wales Public Libraries Division newsletter 7 1982, pp. 7-8.
- 'The Hellenic catalogue system', Australian library news, 12 (4) 1982, p. 12.
- Holmes, Anne, 'Greek cataloguing: other languages too', Multicultural libraries newsletter 3 (2) 1982, pp. 10-11.
- Miller, R. Bruce, 'Non-Roman scripts and computer terminal development', Information technology and libraries 1 (2) 1982, pp. 143-8.
- Rosengren, Jeanette, 'Technilib', LASIE 8 (5 and 6) 1978, pp. 5-12.
- Wellisch, H., 'Multiscript and multilingual bibliographic control alternatives to romanization', Library resources and technical services 22 (2) 1978, pp. 179-190.

RUN 0001

GREEK NETWORK

UPDATE REPORT

X-100007-1 (cont)

X

06 260 00Sa0εσσαλον'ικηSbEγνατ'ιαSc1979  
 07 300 00Sa172 σ.Sc21 ca.  
 08 650 00SaΕλληνικ'η ποι'ηση, Σ'υγχρονη  
 09 960 10SaKokol'asShX.A.  
 10 962 00SaD'odoka poi'etes  
 11 963 00SaGrock poetry, Modern  
 12 982 00SaCARZSbEAA 889.13 KOK  
 13 999 00Sh1101Sa103077-9Sp4.00SgEAA 889.13 KOK  
 CALL NO: EAA 889.13 KOK

X-100008-A \*ri

1/0042

NEW RECORD

CURRENT DATA:

01 001 X-100008-A  
 02 008 s1976SlgrkScgr  
 03 082 00Sa324.249507  
 04 100 10SaΓιανν'ατοςShΣπ'υρος  
 05 245 10Sa0 Γε'ργιος Παπανδρ'εου στ'ις κρ'ισιμες 'ωρες  
 της Ελλ'ας και της δημοκρατ'ιαςSd του Σπ. Γιανν  
 'ατου  
 06 260 00Sa Αθ'ηνα Sb'Εκδοσεις Δ.Καραμπ'ελα-Σγο'υρδαSc197  
 6  
 07 300 00Sa191 σ.Sc21 ca.  
 08 600 10SaΠαπανδρ'εουShΓε'ργιος  
 09 650 00SaΠολιτικο'ιSzΕλλ'ας, Σ'υγχρονηSxBιογραφ'ια  
 10 960 10SaGiannalosShSmyros  
 11 962 10Sa0 Γε'ργιος Παπανδρ'εου ...  
 12 963 00SaStaloscaShSgrecce, ModernSxBiography  
 13 982 00SaCARZSbEAA 324.249507 ΓΙΑ  
 14 999 00Sh1101Sa103076-0Sp7.00SgEAA 324.2495 07 ΓΙΑ  
 15 999 00Sh1102Sa103075-2Sp7.00SgEAA 324.2495 07 ΓΙΑ  
 CALL NO: EAA 324.2495 07 ΓΙΑ

X

X-100009-B \*ri

1/0047

NEW RECORD

CURRENT DATA:

01 001 X-100009-B  
 02 008 s1973SlgrkScgr  
 03 082 00Sa549.507  
 04 100 10Saεσοτα'αςShΓι'ργιος  
 05 245 10SaΕλε'υθρα πνε'υμαSfεπιμ'ελεια: Κ.Θ. Δημαρ'ας  
 06 260 00SaR3'ηναSbEρω'ηςSc1973  
 07 300 00Sa79 σ.Sc17 ca.  
 08 450 00SaH'ca Ελληνικ'η Βιβλιοθ'ηκη  
 09 651 00SaΕλλ'ας, Σ'υγχρονηSxΔιανοητικ'η ζω'η  
 10 700 10SaΔημαρ'αςShΚουσταντ'ινος Θ.  
 11 960 10SaThaslokasShGi'orgos  
 12 962 10SaElouthora pnuoca  
 13 964 00SaGrecce, ModernSxIntellectual life  
 14 982 00SaCARZSbEAA 549.507 ΕΞΟ  
 15 999 00Sh1101Sa040921-9Sp5.00SgEAA 549.507 ΘΕΟ  
 16 999 00Sh1102Sa103074-4Sp5.00SgEAA 549.507 ΘΕΟ  
 CALL NO: EAA 549.507 ΕΞΟ

X

ΕΛΛ ΕΡΓ 489.332 ΠΕΤ: Πετροβ'ιδης, Π: Σύγχρονον αγγλοελληνικόν κα'ι ελληνοαγγλικόν λεξικόν: Χ-800018-2: τμήτ  
0411

ΕΛΛ ΕΥΚ ΛΟΙ: Λο'ιζου, Μ'αρω: Ήταν και δεν Ήταν: 2 εκ.: Χ-800031-κ: 5.00, 342427-8, 0411

ΕΛΛ ΕΥΚ ΠΕΤ: Πεττ, Τσ'ενθα: Τα περιστέρια: Χ-800032-8: 5.00, 391945-5, 0411

ΕΛΛ ΕΥΚ ΧΑΘ: Χαθ'ορν, Λ'ιμπυ: Τ'ο δ'έντρο το'υ Στ'εφανου: Χ-800024-7: 6.26, 305282-6, 0411

ΕΛΛ ΜΥΘ ΒΑΣ: Βασιλικ'ος, Βασ'ιλης: Κρο'υπ-Ελλ'ας: Χ-800001-8: 4.00, 265438-5, 0411

ΕΛΛ ΜΥΘ ΠΑΠ: Παπαδιαμ'αντης, Αλ'εξανδρος: Βαρδι'ανος στ'α σπ'ορκα: Χ-800010-7: 4.00, 265454-7, 0411

ΕΛΛ ΜΥΘ ΣΚΑ: Σκαρ'ιμπας, Γι'αννης: Τυφλοβδομ'αδα στ'η Χαλκ'ιδα: Χ-800000-Α: 4.00, 269612-6, 0411

ΕΛΛ Π ΧΑΡ: Καλι'οτσος, Πανελ'ης: Τ'α ξ'υλινα σπαθι'α: 5 εκ.: Χ-800021-2: 2.50, 353100-7, 0411

ΕΛΛ ΠΛΗ 910.0949 5 ΔΗΜ: Δημητρ'ακος, Σπ'υρος: Δημητρ'ακου γεωγραφ'ια 'ατλας παγκ'οσμιος Ελλ'ας: Χ-800015-8: το

ΕΛΛ ΠΛΗ 910.0949 5 ΔΗΜ: Δημητρ'ακος, Σπ'υρος: Δημητρ'ακου γεωγραφ'ια 'ατλας παγκ'οσμιος Ελλ'ας: Χ-800015-8: το

ΕΛΛ Θ ΛΟΙ: Λο'ιζου, Μ'αρω: Ήταν κα'ι δ'έν Ήταν: 2 εκ.: Χ-800014-Α: 5.00, 342427-8, 0411

ΕΛΛ ΘΩΝ ΤΑΙ ΜΑΡ: Μαριν'ελλα: Μαριν'ελλα σ'αγαπ'ω: Χ-800027-1: 10.35, 384024-7, 0411

ΕΛΛ ΧΑΡ: Ρ'ομπινς, Ντ'ενις: Η ραφ'οδ'ια το'υ 'ερωτα: Χ-800023-9: 4.00, 269572-3, 0411

ΕΛΛ ΧΑΡ: Χριστοδο'υλου, Δημ'ηρης: Τ'ο πατ'αρι: 2 εκ.: Χ-800022-0: 10.00, 353105-8, 0411

ΕΛΛ 181.45 ΜΑΚ: Μακαρτν'ε'υ, Τζ'εημς: Γι'ογκα: Χ-800007-7: 10.00, 372495-6, 0411

ΕΛΛ 291.13 ΔΕΛ: Δ'ελτα-Παπαδοπο'υλου, Αλεξ'ανδρα: Η'υθοι κα'ι θρ'υλοι: 4 εκ.: Χ-800006-9: 4.00, 265591-8, 0411

ΕΛΛ 365 ΣΑΡ: Σαρι'ερ, 'Ανρι: Μπ'ανκο: Χ-800019-0: 22.50, 354575-Α, 0411

ΕΛΛ 365 ΣΑΡ: Σαρι'ερ, 'Ανρι: Ο πεταλο'υδας: Χ-800020-4: 23.00, 370484-Α, 0411

ΕΛΛ 954.045 ΓΚΑ: Ντρ'ημπεργκ, Τρ'εβορ: Ιντ'ιρα Γκ'αντι: Χ-800004-2: 3.50, 120288-Α, 0411

ΕΛΛ ΕΡΓ 489.3321 KPI  
 ΚΡ'ΙΧΤΟΝ ΓΟΥ'ΙΛΛΙ  
 Η'ΕΓΑ ΕΛΛΗΝ'Ο-ΑΓΓΛΙΚ'ΟΝ ΛΕΞΙΚ'  
 ACC NO 103082-5  
 RECORD NO X-100001-2 \$10.00  
 LIBRARY 11 BRANCH 01 14/07/82

ΕΛΛ 889.09 HIP  
 ΗΙΡΑΣΓ'ΕΖΗ ΚΑΡ'ΙΑ  
 ΝΕΟΕΛΛΗΝΙΚ'Η ΛΟΓΟΤΕΧΝ'ΙΑ  
 ACC NO 103072-8  
 RECORD NO X-100011-A \$15.00  
 LIBRARY 11 BRANCH 01 14/07/82

ΕΛΛ Π 398.2094 5 ΝΕΓ  
 ΝΕΓΡΕΠ'ΟΝΤΗΣ ΓΙ'ΑΝΝΗΣ  
 15 ΔΕΚΑ-ΠΕΝΤΕ ΠΑΡΑΜ'ΥΘΙΑ  
 ACC NO 103066-3  
 RECORD NO X-100016-0 \$6.00  
 LIBRARY 11 BRANCH 02 14/07/82

ΕΛΛ Π 398.2094 5 ΝΕΓ  
 ΝΕΓΡΕΠ'ΟΝΤΗΣ ΓΙ'ΑΝΝΗΣ  
 15 ΔΕΚΑ-ΠΕΝΤΕ ΠΑΡΑΜ'ΥΘΙΑ  
 ACC NO 103067-1  
 RECORD NO X-100016-0 \$6.00  
 LIBRARY 11 BRANCH 01 14/07/82

ΕΛΛ 949.507 ΘΕΟ  
 ΘΕΟΤΟΚ'ΑΣ ΓΙ'ΩΡΓΟΣ  
 ΕΛΕ'ΥΘΕΡΟ ΠΝΕ'ΥΜΑ  
 ACC NO 103074-4  
 RECORD NO X-100009-8 \$5.00  
 LIBRARY 11 BRANCH 02 14/07/82

ΕΛΛ 949.507 ΘΕΟ  
 ΘΕΟΤΟΚ'ΑΣ ΓΙ'ΩΡΓΟΣ  
 ΕΛΕ'ΥΘΕΡΟ ΠΝΕ'ΥΜΑ  
 ACC NO 040921-9  
 RECORD NO X-100009-8 \$5.00  
 LIBRARY 11 BRANCH 01 14/07/82





1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

66-MCULTP-6-E  
INF

Character Set and its Control in an Online Cataloguing System

Application at the Center for Bibliographic Information

by

Akira MIYAZAWA  
Tokyo, JAPAN

KEYWORD(S):

For internal use only:

To be brought to meeting no:

Friday 22 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

## 1.はじめに

東京大学文献情報センター\*は、日本の大学図書館に対してオンライン共同目録システム(CBIシステム)を提供している機関である。この目録システムには、基本として大学図書館の購入する資料に使われるすべての言語データを収容するデータベース・システムを構築することが求められていた。

CBIシステムは、書誌ユティリティ・システムとして次のような新しい特徴を持っている。1) データベース管理システムとコンピュータ・ネットワーク・システムに基づいた新しい設計。2) 分散処理志向。3) 総合目録。4) 統合化された典拠コントロール・システム。5) インテグリティ・チェックのためのソフトウェア・システム。6) 画面型インタフェース。7) 漢字、ALA文字セットのサポート。

また、アルファベットと漢字の入力機能に加え、ALA文字を入出力するために、目録端末が開発された。

## 2.文字セット

1つのデータベース・システムで使われる文字コード・テーブルは「文字セット」を形成する。多文字種を処理する際に最初に問題となるのは、そのデータベースに対して十分な文字セットを作成することであり、次に問題となるのは、その文字セットを扱うデバイスを手入手することである。

CBIシステムはN-IXNVT(拡張仮想端末)のコードテーブルを「文字セット」としている。表-1は現在(1986年3月)CBI文字セットにおける文字種と数を示している。表-1の示すように、現在のCBIの文字セットは、日本語、英語、ドイツ語、フランス語、その他のローマ字を使用する言語、さらにロシア語、ギリシャ語をカバーしているが、中国語、韓国・朝鮮語その他のアジア諸語はカバーしていない。

この文字セットでは、1バイト2バイトの文字の混合文字列を取り扱う必要がある。この処理のため、'mixed character'型の概念を導入し、この型の文字列を扱うルーチン群がシステムに用意されている。

## 3.文字セット管理システム

漢字のように5万を越える文字種を持った文字セットでは、そのすべてを数えあげることは殆ど不可能である。「開いた」文字セットシステムとは、固定された数の字種を持つものではなく、漢字のような場合、必要に応じて字種を増やすことのできるものである。「開いた」文字セットを管理する文字セット管理システムは、次のような3つの機能を持つ。

- 1) システム外字の登録機能;
- 2) 入力外字の入力機能;
- 3) 出力外字の出力機能。

CBIシステムでは、システム外字の登録機能は集中化されており、文献情報センターにおいて実行される。

また、外字は特別なフォーマットで入力され、システムはこれを見て外字処出現の報告を行なう。文献情報センター内の一部所がこれを集め、必要に応じて文字の登録を行なう。

入力の困難を避けるため、ユーザは入力外字を16進コードで入力することができる。また、いくつかの漢和辞典の検字番号を用いて入力することも可能である。

## 4.将来の課題

現在のCBIシステムの文字セットは、中国語、韓国・朝鮮語やアジア諸言語をカバーしていない。この点は近い将来大きな問題となりうるであろう。また現在の目録端末では不可能なキリル文字の直接入力も望まれている。さらに多言語環境でのキーワード切出しなど、言語処理機能の研究開発を進めて行く考えである。

\* 1986年4月から東京大学文献情報センターは、学術情報センターとなったが、混乱を避けるため、本稿では旧称を使用する。

# Character Set and its Control in an Online Cataloging System

## Applications at the Center for Bibliographic Information.

Akira MIYAZAWA

National Center for Science Information System.

### Abstract

This paper describes the current status of the multiscrypt processing of a cataloging system, which is operated in the Center for Bibliographic Information University of Tokyo. The CBI system is the first nation wide 'bibliographic utility system' in Japan, and based on database and computer communication technologies. The features of the CBI system are described. The character set of the system includes alphabets, *kanji*, *kana* for Japanese and also alphabets with diacritical marks. Then method of extension of the character set is described. Finally, problems in current stage and future prospections are discussed.

### 1. Introduction

Center for Bibliographic Information University of Tokyo(CBI)<sup>1</sup> services an online shared cataloging system for Japanese university libraries(CBI system). In this system, the basic requirement was to develop a database system for bibliographic data of all the languages used in materials acquired by university libraries.

Japanese computer technology for processing Japanese language has been rapidly advanced in recent years. Based on this technology, the CBI system realized multiscrypt database system for library materials.

---

<sup>1</sup> In April 1986, this center was reorganized as the National Center for Science Information System. However, to avoid confusion, this paper refers the old name.

This paper reports the current stage of the CBI system and its character set handling.

## **2. CBI cataloging system**

### **2.1. Center for Bibliographic Information**

The Center for Bibliographic Information, University of Tokyo was established in April 1983. One of CBI's mission is to develop and operate so called 'bibliographic utility' system for Japanese university libraries.

A computer system, HITAC M-280H, was installed in the end of February 1984. Since then software development of new cataloging system were started. CBI system is based on a database management system Model 204, and N-1 network system.[1]

### **2.2. Features of CBI cataloging system**

The CBI system has several new features amongst bibliographic utility systems. They are

- (1) Up-to-date design with database management system and network system : this system is base on structured database design and computer network system(N-1).
- (2) Direction for distributed processing system : this system is not comprehensive one for total library automation. However, linking with in-house systems of each libraries, it will become a core system of the science information network.
- (3) Union catalog : this system is strongly directed for union catalog database production, which will be a base of new science information system for university researchers.
- (4) Integrated authority control system : in the database of this system authority records and bibliographic records are liked each other, and an efficient linking mechanism is provided by the system.

- (5) Software system for integrity check : link integrity and unification of bibliographic data are supported by software systems.
- (6) Screen oriented application : a user friendly screen type interface is provided. In a computer network on different manufacturer's host system, screen type interface is distinctive.
- (7) Kanji and ALA character set support : letters with diacritical marks -- such as umlaut, acute -- are utilized as well as kanji, and kana characters for Japanese language.

Amongst these features, the last one is fully discussed in the latter part of this paper.

### 2.3. Cataloging workstation

Cataloging workstation is specially designed to manipulate both ALA characters and *kanji* characters.

Cataloging workstation manufactured by HITACHI is based on an intelligent terminal T560-20 PS. Original T560-20 PS intelligent terminal has Japanese character input capability by *kana* to *kanji* conversion mechanism. The cataloging workstation has ALA character input/display capability, in addition to original input capabilities. Table 1 shows the input repertory of the cataloging work station.

Diacritical marks are labeled on the front side of keys. And a special shift key(non lock) called 'FUN' is used to input diacritical marks. Normal alphabets following after diacritical marks are immediately composed into one letter with the previous diacritical mark.

However, not all the compositions are allowed. There are just 685 composed character set and the workstation hardware logic selects only valid compositions.



**Table 1. Input repertory of the cataloging work station**

| Type                                   | Example | Display Width | Keying Method                                    | No. of Characters |
|----------------------------------------|---------|---------------|--------------------------------------------------|-------------------|
| Roman Alphabets                        | a b c   | Half Width    | Direct Keying                                    | 158               |
| Alphabet Symbols                       | , : #   |               |                                                  |                   |
| Kana                                   | ア イ ウ   |               |                                                  |                   |
| Kana Symbols                           | ・ ・ ・   |               |                                                  |                   |
| Kanji                                  | 重 以 宇   | Full Width    | Code Input                                       | 8877              |
| Kanji Symbols                          | ☆ ◎ ※   |               |                                                  |                   |
| Greek Alphabets                        | α β γ   |               |                                                  |                   |
| Cyrillic Alphabets                     | а б в   |               |                                                  |                   |
| Special Alphabets                      | æ ð ð   | Half Width    | Direct Input                                     | 885               |
| Alphabet with Diacritical Marks        | á â ç   |               | Multiple Keying of Diacritical mark and alphabet |                   |
| Alphabet with Double Diacritical Marks | ä ö ø   |               |                                                  |                   |
| Isolated Diacritical Marks             | · · ·   |               | Direct Input used temporarily                    |                   |

### 3. Character set

#### 3.1. Character code and character set

Generally, character code assignments for a graphic character vary according to each input/output devices. However, in a database system, which is used by many users, character code and graphic character assignment must be fixed uniquely.

Characters in this code assignment table forms a 'character set'. The first problem of multiscript processing is to make sufficient character set for the databases. And second is to obtain input/output devices to handle its character set.

#### 3.2. Character set of CBI system

The CBI system use N-1 XNVT(eXtended Network Virtual Terminal) code table[2] as its 'character set'.<sup>2</sup> N-1 XNVT code is composed of characters in JIS C8220, JIS

<sup>2</sup> Internally, N-1 XNVT code is mapped on EBCDIK and KEIS code.

C6226 and Extended character set(EXC) as its 'character set'.

JIS C6220 is 1 byte code table with ISO 646 and Japanese *kana* characters. JIS C6226 is 2 bytes(7 bits) code table for Japanese *kanji* characters. These two character set is Japanese standard, and most of Japanese systems have capability of handling these two character sets.

EXC is a 2 bytes(8 bits) code table special for the CBI system. This character set is an open-end character set, which may be extended according to database requirement.

Currently, 691 characters including special alphabets(Polish L, etc.), alphabets with diacritical mark(umlaut o, etc.), and some special symbols are defined as EXC character set.[3]

Table 2 shows types and numbers of characters in current(March '86) CBI character set.

As Table 2 shows, present character set of the CBI system covers Japanese, English, German, French, and other languages using Roman alphabets, and Russian, Greek as well, but not for Chinese, Korean and other Asiatic languages.

### 3.3. Multi code space processing

N-1 XNVT code is composed of 2 code spaces. One is 1 byte(8 bits) and the other is 2 bytes(8 bits) code space. Because of this structure, software development had to take care of this multi code space character set processing. One is technic on mixed space string processing, the other is duplicity of characters in two spaces.

Processing of the strings which contains both 1 byte and 2 bytes characters have to be taken care. For this processing, the concept of mixed character type[4] is introduced, and handling routines for mixed character strings are installed on the system.

**Table 2. CBI character set**

| Type                             | Example                   | No. of Characters | Code Space | Note      |
|----------------------------------|---------------------------|-------------------|------------|-----------|
| Roman Alphabets                  | a b c                     | 52                | 1 byte     | JIS C6220 |
| Digits                           | 0 1 2                     | 10                |            |           |
| Alphabet Symbols                 | . : #                     | 31                |            |           |
| Kana                             | ア イ ウ                     | 55                |            |           |
| Kana Symbols                     | ・ ・ ・                     | 8                 |            |           |
| Kanji Symbols                    | ☆ ◎ ※                     | 147               | 2 bytes    | JIS C6226 |
| Digits                           | 0 1 2                     | 10                |            |           |
| Roman Alphabets                  | a b c                     | 52                |            |           |
| Hira-kana                        | あ い う                     | 83                |            |           |
| Kata-kana                        | ア イ ウ                     | 86                |            |           |
| Greek Alphabets                  | $\alpha$ $\beta$ $\gamma$ | 48                |            |           |
| Cyrillic Alphabets               | а б в                     | 66                |            |           |
| Lines                            | └ ┘ ┐                     | 32                |            |           |
| Kanji                            | 亜 以 宇                     | 6353              |            |           |
| Symbols                          | - . ~                     | 30                |            |           |
| Special Alphabets                | æ ð ß                     | 23                |            | EXC       |
| Alphabets with Diacritical Marks | á â ç                     | 632               |            |           |
| Graphics for Control Characters  | ☐ ☑ ☒                     | 6                 |            |           |
|                                  |                           |                   |            |           |

March 1986.

Mixed type string may be a string of 1 byte and/or 2 bytes characters. On external devices such as terminals, mixed type strings are represented with shift codes to/from 1 byte from/to 2 bytes. In application programs, they are represented by all 2 bytes character string, padding *null* character before 1 byte character. This conversion routine, and some basic manipulation routines on mixed type strings are developed. In the CBI system, all the character string manipulations are done by this mixed type string.

By this mechanism, character set processing of multi code space is realized consistently in the CBI system.

Besides this, the duplicity of characters may cause difficulties. For example, Roman alphabet 'A'(capital) has code  $(41)_{16}$  in 1 byte space, and  $(2341)_{16}$  in 2 bytes space. In some systems, this situation causes a confusion in keyword searching. In our system a convention is made to avoid this confusion. That is :

A character, which is included both in 1 byte and 2 bytes space, is stored as 1 byte code.

Users may input in either of code(i.e.  $(41)_{16}$  or  $(2341)_{16}$  for 'A'), and it is stored(or searched) as 1 byte code(i.e.  $(41)_{16}$ ).

Therefore a user of our system may be unconscious about the code which his terminal outputs.

#### 4. Character set control system

##### 4.1. Open-end character set

In the characters as *kanji* of which number may exceed fifty thousands, it is almost impossible to count up all of them. For this reason, I proposed a concept of 'open-end character set system' for the processing of large character set.[5] An open-end character set system is not fixed in the number of characters, and allows the number of characters in the character set to be increased, as necessary.

For materizing an open-end character set system, so called external character handling function is required for both hardware and software system. In Japan, recent development of *kanji* processing allows full width of selection for such hardware and software system.

##### 4.2. Character set control system

In a open-end character set system, to control character set, two factors is required. they are: 1) a method of assigning a new character into the character set, and 2) a mechanism of input and output the new assigned character. A system which includes these two factors shall be referred to as character set control system. Usually a character set control system has next three functions.

- 1) Registration function of the system external character;
- 2) Input function of the input-external character;
- 3) Output function of the output-external character.

Character set control systems are employed in some database systems in Japan, such as in National Institute of Japanese Literature.[6]

#### 4.3. Character set control in CBI system

In the CBI system, the registration function of system external characters is centralized, and performed by the CBI. External characters are input in special format, and then the system reports the existence of external characters. A section in the CBI collects the reports and register them, if necessary.

The CBI system is a network system on various computer systems. This situation makes difficulties about input and output functions of external characters, because input/output devices vary for each users terminal system.

To avoid input difficulty, the CBI system allows user to input external characters by hexadecimal code. In addition, they can also input them by the reference number of the character in some dictionaries.<sup>3</sup>

Output functions are available in some user terminal systems and cataloging workstations. The central system assumes that user terminal have facilities to output external characters. User terminal(or system) may neglect their output-external characters and display them as special marks to show it is an output-external character. This limitation is not critical for Japanese language and current character set.

#### 5. Future prospects

As mentioned in previous chapter, current character set of the CBI system does not cover Chinese, Korean and other Asiatic languages. *Kanji* characters in JIS C6228 is not sufficient for Chinese.

---

<sup>3</sup> such as Morohashi's 'Dai kanwa jiten'.



Open-end character set system allows character set to be increased. However, for the Chinese language, the number of characters to be added may be ten to forty thousands. Theoretically, the number less than thirty thousands may be acceptable. Yet practically, ten thousand causes serious difficulties, for hardware and software systems. Solution of this problem will be important in near future.

Next problem is insufficiency of Cyrillic input devices. As shown in Table 1, the cataloging workstation can not input Cyrillic alphabets directly. To input them, code input method or conversion input method, which takes three to four key touch for one character, is available. These methods are not sufficient for catalog input, and direct keying terminal is being desired.

I did not mention about language processing such as keyword extraction, in this paper. Currently, these functions are implemented only in very simple algorithms. However, this will be the next stage of our research and development, especially in multilingual circumstances.

#### ACKNOWLEDGEMENT

I am grateful for Mr. Eisuke Naito for his advise and encouragement.

#### REFERENCES

- [1] NAITO, Eisuke, MIYAZAWA, Akira and INOUE, Hiroshi. Library networking in Japan : the role of the Center for Bibliographic Information.
- [2] ADACHI, Jun. Enhancement of N-1 network virtual terminal protocol. Research Bulletin of the Center for Bibliographic Information, University of Tokyo. Vol 1.(1985), p.130-135.[in Japanese]
- [3] MIYAZAWA, Akira. Extended character set at the CBI system : Selection and implementation. Research Bulletin of the Center for Bibliographic Information, University

- [4] SHIMAZAKI, Masaki, HAYASHI, Makoto, TSUDA, Takao. PLAK/R : Programming Language with Abstract data type for Kanji Processing and Relational database operations. Proceedings of ICTP'83(1983), p241-246.
- [5] MIYAZAWA, Akira, TAJIMA, Kazuo. Character set control system and 'Kanji thesaurus'. Proceedings of ICTP'83(1983), p379-386.
- [6] National Institute of Japanese Literature, Information Processing Section, ed. Kanji character set for data processing system. NIJL Technical report, No. 8. (1981).

IFLA

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

1986 Tokyo, August 21-22

67-MCULTP-7-E  
INF

UTLAS DEVELOPMENTS WITH JAPANESE AND CHINESE SCRIPTS

by:

Jack CAIN  
Advisory Analyst  
Utlas International Canada, a division of International Thomson Limited  
80 Bloor Street West, 2nd Floor  
Toronto, Canada  
M5S 2V1

KEYWORD(S):

Japanese script  
Chinese Ideograms  
Online retrieval  
Utlas

For internal use only:

To be brought to meeting no:

Friday 22 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

### 1. Utlas International Canada

Utlasは図書館および情報産業を対象とするコンピュータ資源活用のサービス機関である。Utlasは書誌情報提供機関として、1973年よりオンライン・データベース・サービスおよび関連製品を英語そして仏語で提供してきている。カナダ、アメリカそして日本で、コンソーシア、政府機関の参加メンバーを含めた465をこえる機関、図書館において2,000以上の参加館がUtlasの機能を通じて、利用者自身のデータベースの構築・維持を行なっている。Utlasは、また、最新のコンピュータ技術を生かした図書館内システムの提供も行なっている。日本では、丸善株式会社がUtlasの代理店となっている。

Utlasにおける多文字処理システムの開発活動について、また、多文字処理の基本的な設計を決定することにより出すに到るまでの調査研究に関して述べる。

多文字処理にアプローチするにあたり、Utlasは新しいサービスの開発に努力しており、その国の伝統、言語、および文化を理解し、尊重していきたいと考えている。このようなアプローチを試みるのは、その国の要請にできる限り応えられるシステムを作り上げるためである。

### 2. オンライン多文字処理システム開発

環境(各国の伝統、言語、および文化)は、その国の必要とする書誌調整の形式に、基本的なレベルで影響を及ぼしている。このことを考えにおくとアジアで、また他のどこでも、別々の文字の、異なる目録の慣習というもの、その環境の非常に大きな部分となることは明らかである。アジアにおいてもそれは顕著であって、中国語と日本語では実用的な目的から、二つの別々の文字と異なる目録が一般に用いられている。

### 3. 多文字の機械化処理のための設計構想

a)費用; b)利用者共同体; c)文字セット; d)通信; e)入力方法論; f)ユーザーインターフェース; g)MARC(機械可読目録)形式; h)アクセス手順;

### 4. Japan CATSS(Catalogue Support Service)

日本の現在の利用者の要求に見あうことを考慮してUtlasはJapan CATSSのプロトタイプ(ひな型)を開発している。従来CATSSは目録作成および関連業務のためにUtlasの利用者によって用いられているローマ字だけのオンライン共有目録作成機能のことである。プロトタイプを作成するために、現行のWestern CATSSシステムが複製され、日本語での操作ができるように改変されている。プロトタイプは以下のような特徴をそなえている。

- a) アジアの伝統に従って、Japan CATSSは日本語のためだけの個別のシステムとする;
- b) 文字セット: JIS(C6226);
- c) Japan MARC フォーマット;
- d) "カナ漢字変換"入力手順;
- e) オンライン会話型-日本語のメッセージおよびプロンプト;
- f) アクセス機能(-ISBN, 他の番号による数値検索キー; -著者名、書名、叢書名および件名による拾読通覧キー; -論理演算検索; -カナ検索法-カナと漢字の表示;)

### 5. 中国語および韓国語の開発計画

- a) 各々、個別のシステムとして計画を持っている。
- b) 文字セットおよびMARCフォーマットは、その国の標準に準拠する。
- c) 韓国語の入力はハングルで、中国語は合成方式で行なう。
- d) 韓国語の拾読通覧キーはハングル順配列とし、中国語の拾読通覧キーは、発音順と字画順配列の二種類とする。

Utlas is a computer-based service organization for libraries and the information industry. In this paper, a description is given of Utlas' online cataloguing system which operates in Japanese. The basic design principles and the chief functional characteristics of the Japanese system as well as Utlas' plans for Chinese and Korean online processing are provided. A fundamental aspect of Utlas' approach to this development is that it is essential to mold the design of the automated system to fit the needs of the library patrons while respecting the "environment" of the country concerned - its traditions, language and culture.



## CONTENTS

1. Introduction
2. Bilingual Experience
  - a) Online Interactive Dialogue
  - b) Bilingual Records
  - c) Bilingual Printed Products
  - d) Bilingual Access
  - e) Bilingual Authorities and Automatic "Translation"
3. Online Multiscript Development - Research
  - a) Library Patron Needs
  - b) Romanization and Transliteration
  - c) Direct Access to Ideographic Characters
4. Design Considerations for Automated Multiscript Processing
  - a) Costs
  - b) User Community
  - c) Character Set
  - d) Telecommunications
  - e) Input Methodologies
  - f) User Interface
  - g) MARC formats
  - h) Access Mechanisms

5. Japan CATSS

- a) Separate System
- b) Character Set
- c) MARC Format
- d) Input Methodology
- e) User Interface (Online Interactive Dialogue)
- f) Access Mechanisms

6. Plans for Chinese and Korean

- a) Separate System
- b) Character Set
- c) MARC Format
- d) Input Methodology
- e) User Interface (Online Interactive Dialogue)
- f) Access Mechanisms

7. Conclusion

## 1. Introduction

Utlas is a computer-based service organization for libraries and the information industry. As a bibliographic and information utility, Utlas has been supplying online database services and related products in English and French to customers since 1973. Over 475 institutions, members of consortia and government agencies, representing over 2,000 libraries in Canada, the United States and Japan, maintain user-owned databases through Utlas facilities. Utlas also provides local library systems based on state-of-the-art computer technology.

Utlas is represented in Japan by the Maruzen Company, Limited.

In this paper, the author presents the activities leading up to multiscript development at Utlas as well as the research which was undertaken to arrive at basic design decisions for multiscript processing. In its approach to this processing, Utlas has undertaken to respect the traditions, languages and cultures of the countries for which it is striving to develop new services. Thus, for example, Utlas has chosen, for Japan, to develop a separate system exclusively for the use of Japanese libraries in processing Japanese publications. This approach has allowed the system to be tailored as much as possible to national needs.

## 2. Bilingual Experience

Since it began in a Canadian environment, Utlas has always offered bilingual services - in English and French. With more than a decade of experience in processing two languages, Utlas was well prepared in 1984 to begin preparing for other languages and scripts.

The bilingual, English/French, service affects five areas of Utlas services:

a) Online Interactive Dialogue

In interaction with the online system, the user has always been able to receive messages, prompts and give commands in either language as well as being able to switch back and forth from one language to the other quickly and conveniently.

b) Bilingual Records

Early on in the development of the Canadian MARC format<sup>1</sup>, provision was made to allow both English and French versions of most bibliographic elements to reside in parallel fields. Utlas file structures (which allow each user to build his own database of records) also allow users to build records with the CAN/MARC style parallel fields. In addition, users are also allowed to build parallel files, one for English records, one for French records.

c) Bilingual Printed Products

From the beginning of its services, Utlas offered all products in both languages - catalogue cards, book catalogues, COM (Computer Output Microfiche and Microfilm) catalogues, acquisition lists, labels, and MARC tapes.

d) Bilingual Access

Since users can create headings in either language, the online retrieval software provides access in both languages.

e) Bilingual Authorities

Just as users can create their own databases of bibliographic records, they can also create their own databases of authority records and link them to the corresponding language headings in bibliographic records. The online software keeps track of which language is which and only links French authority records to French bibliographic headings, and English authority records to English bibliographic headings.

### 3. Online Multiscript Development - Research

Before beginning online multiscript development, Utlas engaged in considerable investigation into the manual practices used by libraries to control multiscript bibliographic data. This investigation was not limited to practices in North America since, at that time, Utlas already had user libraries in Asia.

#### a) Library Patron Needs and the Library Environment

In our investigation, it was clear almost immediately, that the types of bibliographic control provided in manual record systems are determined to a large extent by the needs of the majority of the patrons of the library in question as well as the "environment" ( the traditions, language and culture of the country) of that library. The effort to respect the library patron needs as well as the special features of the library's environment led us to conclude that these considerations must be fundamental to our approach to developing systems for other countries, especially for countries with other than Roman script languages.

In the Americas, where a single script - Latin - has been overwhelmingly predominant, the manual catalogue records representing items in other scripts are fully or partially in Latin script. Since many users of the "other script" material are more familiar with Latin script than the script of the material they are using, this seems to be a reasonable expedient. Those for whom the "other script" is more familiar are usually a small minority.

In China, Japan, or Korea, however, Latin script (although increasingly familiar) is not more familiar than the local script. Library patrons therefore are not assisted by having Latin script imposed on the local script.

Given this local environment for libraries, it is not surprising that a system of separate catalogues for separate scripts is the norm in China

and Japan. It is significant too that in both these countries Chinese and Japanese are usually considered, for practical purposes, to be two separate scripts and separate catalogues are maintained for them.

b) Romanization (Conversion into Roman or Latin letters)

In the Americas, a single catalogue for all scripts is the norm. (Harvard being a notable exception.) This single integrated catalogue is achieved by the romanization of all access points so that the filing of entries may be done entirely according to filing rules that need account only for the 26 "Latin" letters used in English.

In China, the promotion of pinyin romanization of the Chinese script has led to the creation of many catalogues where some or all access points are filed under pinyin. In spite of this practice, integrated catalogues are not the norm; separate catalogues are usually found for Chinese, Japanese, Russian and Western catalogue records.

In Japan, separate catalogues are also common. Of course the Roman script is well known in Japan and some catalogues are filed under Romanized forms. The Japanese language, however, since it employs in its modern written forms both syllabic characters (katakana and hiragana) as well as ideographic characters, usually of Chinese origin, already has the possibility of using its own syllabic characters instead of Roman letters to arrange catalogue entries. This is of course what happens in many cases. In some cases, catalogue entries are filed by pronunciation without the syllabic representation being explicitly present on the card; characters whose pronunciation is obscure may be added in pencil on the verso of the card to aid the filer.

c) "Direct" Access to Ideographic Characters

In China, it is not uncommon to find manual catalogues which are provided with "direct" access to the ideographic characters used to write the language. By direct access is meant a method not based on pronunciation. Examples of these methods are a) total strokes in the



character, b) character radical plus additional number of strokes, c) stroke type subfiled by total strokes, and d) total strokes subfiled by stroke type.

Catalogues which use the "direct" access method avoid the difficulties associated with catalogues which use romanization or other systems based on pronunciation. In China, two such difficulties are regional accents and the problem of the high occurrence of homonyms. Direct access avoids both these problems although it has other difficulties, the chief of which is, perhaps, that it rarely provides a unique sort sequence; however, native speakers are used to this short-coming in their dictionary arrangements. A good example of the utility of such catalogues is the case of library catalogues in Hong Kong where romanized access does exist, but it is based on the northern pronunciation not Cantonese. The result is that the direct access catalogues are much more heavily used because they are more convenient to native Hong Kong patrons.

These same principles apply to catalogues of Chinese material found in Japan.

#### 4. Design Considerations for Automated Multiscript Processing

##### a) Costs

A basic consideration, in the development of multiscript processing as in all other Utlas development, is to achieve a reasonable balance between cost and the provision of services. Cost is affected both by the complexity of functionality and the efficiency of its implementation. Therefore, it is important to choose to implement only a level of complexity that libraries will be able to afford and to implement it as efficiently as possible. Given the complexity of Asian languages, this particular design consideration is all the more important.

b) User Community

In designing multiscrypt systems, the second most important criterion after cost is that of the user community. Who will be using the system? What are the needs of these patrons?

In studying the situation of libraries in China and Japan, it is clear that these two distinct library communities, because of the differences between their linguistically unrelated languages, require major differences in the design of library automation systems destined for their use.

c) Character Set(s)

In considering the automation of any multiscrypt processing, the question of character set or sets is of course the first to arise. With languages which include Chinese ideograms - Chinese, Japanese, Korean and formerly Vietnamese, the question of character set is much more difficult than with any other group of languages. The most important areas of difficulty are:

i) Size

Between 70,000 and 100,000 unique characters are known to exist, although the majority of them are rare.

ii) Open-Endedness

In any application it is not possible to predict exactly which characters will be needed, and this consideration is more important with bibliographic data than with other applications. Therefore, provision must be made to add characters to the set.

iii) Standards

Various standards exist. The International Standards Organization in Geneva is working on one, but it will be some years before it is published.

Japan has a national standard (JIS C6226), but it has not been tested, so far as we know, with large amounts of pre-modern bibliographic data. In addition, although JIS is widespread in external communication, many de facto internal standards exist in Japan. Characters outside JIS must be agreed to manually by the exchanging parties concerned.

Several standards are in use both in Taiwan and Korea. In China, one standard has been issued, but appears to be undergoing extensive revision.

In America, the Research Library Information Network, as we have heard, has developed a standard which is international in scope and which follows in its arrangement and structure one of the standards developed in Taiwan.

In assessing these standards, a flaw with all of them (with the possible exception of the new standard being developed in China) is that the set is not in a single sequence that is designed to be used for sorting. It is our opinion that, for efficient retrieval from large databases it will be highly desirable for the set to have this feature. Unfortunately, the sets in existence were designed before retrieval from large (>1 million records) databases came into existence.

d) Telecommunications

In external communication, it is important to conform to standards that will not cause difficulties in the transmission of data. The external standards mentioned above, of course, do conform to the international standards generated by ISO (International Standards Office). Efficiencies of internal processing, however, may dictate that the internal standard is quite different.

e) Input Methodologies

A great deal of work has been done in this area and standards are now beginning to emerge. In choosing input methodologies, however, Utlas decided early on that certain principles would govern its choice of terminal equipment:

- i) Input methods would be chosen which were not experimental, but were firmly established, recognized as efficient, and accepted by the user community.
- ii) Such methods must be implemented on standard equipment which is readily available at reasonable cost, and is being maintained as a commercial product.
- iii) The methods must be designed for use by the "casual" user, and not designed solely for professional input operators.

f) User Interface

The user interface (online interactive dialogue) must of course be able to function in the language of those persons for whom it has been designed.

g) MARC (Machine Readable Cataloguing) Formats

In designing any library system at the present time, a challenging obstacle is the current multiplicity of MARC formats, both internationally and within single countries. It seems inevitable that the systems designer will face the difficult task of building conversion programs to allow data to move, (as cleanly as possible) from one format to another.

h) Access Mechanisms

The design of the access mechanism is of course very much affected by the needs of the user community (see 4 b) above).

In Japan, for example, the principle access to Japanese language data must be through pronunciation. Any system specifically designed for Japanese users must recognize this basic fact.

In China, however, two types of access seem to be necessary, one based on pronunciation, one based on character form.

Any system designed to serve both China and Japan simultaneously will do so only by making sacrifices in design to either or both.

## 5. Japan CATSS (Catalogue Support Service)

In attempting to meet the needs of its current customer base in Japan, Utlas has developed a Japan CATSS system. (CATSS is an online shared cataloguing utility currently used by Utlas customers for cataloguing and related activities in Roman script only.) The methodology used by Utlas was to develop a pilot system (called the Prototype). The Prototype provides a functional system which includes all the basics of a complete system, but is still open for modification based on user response.

In order to create the Prototype, the existing Western CATSS system was replicated and modified to operate in Japanese. A permanent system, located in Japan, is planned to be developed and made available for use. The Prototype has the following characteristics.

### a) Separate System

It was clear to us quite early that only by creating a separate system would it be possible to respond as fully as possible to the needs of the Japanese users for whom the system was designed.

### b) Character Set

In following national standards, the Prototype has adopted the Japan Industrial Standard character set (C6226) in external communication.

c) MARC Format

In following national standards, the Prototype has adopted the Japan MARC format (2nd edition. Tokyo, National Diet Library, 1985). Fields have been added for institution-specific data such as call number, local notes, accession numbers, etc. (See sample in appendix.)

d) Input Methodology

The system uses the kana to kanji conversion input system which is now the "mainstream of development in Japan".<sup>2</sup>

e) User Interface (Online Interactive Dialogue)

The User Interface is entirely in Japanese. As an option, it is possible to switch back and forth with English.

f) Access Mechanisms

Utlas standard access mechanisms consist of two kinds of indexes to the online system:

- i) Precise indexes
- ii) Browsable indexes

Precise indexes are numeric or alphanumeric, such as ISBN (International Standard Book Number).

Browsable indexes, a unique feature of the Utlas system, are indexes which allow the user to "browse" or move backward and forward alphabetically in the keys at the access key level. This "browsing" is very much like moving back and forth alphabetically in a manual catalogue. The browsing may also be combined with boolean operators (and, or, not) to refine the search.



The precise indexes in the Prototype operate just as those do in the Western system, one enters a number and receives a reply showing the record(s) found with that key.

The browsable indexes work the same in principle as the Western browsable indexes, except that they are entirely in Japanese. A search argument is entered in katakana characters, and the system replies with the heading which is the closest match. This heading is presented not in its katakana form, but in the full Japanese form (kana plus kanji, i.e., both syllabic and ideographic characters). The user is then free to "browse" forward or backward in the index. As this browsing takes place, headings are presented, one after the other, filed in the standard order of the katakana syllabary, but presented in full kana plus kanji form. (See sample in Appendix.)

## 6. Plans for Chinese and Korean

### a) Separate System

A separate system is planned for each of Chinese and Korean.

### b) Character Set

The character set used will be based on a standard developed by the user community for whom the system is designed.

### c) MARC Format

The MARC format used will be based on a standard developed by the user community for whom the system is designed.

### d) Input Methodology

The Korean system will be hangul input.

The Chinese system will offer either pronunciation input or the character component method. The choice will be determined by the prospective users.

e) User Interface (Online Interactive Dialogue)

The user interface will be in the language of the user community.

f) Access Mechanisms

The browsable indexes in the Korean system will operate in hangul.

The browsable indexes in the Chinese system will be of two types:

- i) pronunciation order
- ii) character form order

7. Conclusion

It has been about 20 years that librarians and computer professionals have struggled with the automation of library functions. There is some agreement now, at least among those knowledgeable in the field, that this kind of automation is one of the most difficult. Some of the reasons are: the data is truly variable in nature and cannot be modified to suit the computer's natural inclination for well defined spaces. The character set of 94 symbols (in Roman alphabet applications) adequate for most applications, is not enough for libraries. The databases are often very large and retrieval from them is complex. Different library functions represent, in essence, totally different computer applications - the functions of cataloguing, interlibrary loan, acquisitions, printed products, for example, do not resemble each other from a software design point of view. Finally, different types and different sizes of libraries because of differing needs require differently designed systems. Given the difficulty of the task, what has been accomplished so far is remarkable.

In this very specialized (from a computer point of view) field of processing non-Roman scripts, the librarian and the computer professional as a team, once again find themselves at the forefront of the capability of the electronic equipment at their disposal. The available tools - hardware, systems software and programming languages - were not designed with non-Roman processing in mind. All three of these essential components of automation have at their basis the assumption that 94 symbols should be adequate. To go beyond 94 symbols in fact does violence to their fundamental design. The initial response from most computer specialists when you say you're working in the field of multiscript processing is Good Luck!

Perhaps in the not too distant future the above mentioned tools will evolve into something more usable by libraries by including in their foundation room for multiscript processing. Certainly great strides have already been made in the area of handling massive amounts of textual data. In the meantime, those faced with the automation of multiscript records will enjoy a unique and intensely challenging moment in the history of library development.

We hope that the approach taken by Utlas in its initial multiscript processing efforts will be of interest to those present as well as to readers of the published paper. We are pleased and honoured to have had the opportunity to participate in this challenging new aspect of library automation.

#### FOOTNOTES

1. Canadian MARC Communications Format: monographs. 3rd ed. (Ottawa, National Library of Canada, 1979).
2. "Japanese (Sino-Japanese) Characters for Bibliographic Machine Handling; Status Report" [by] Nakamoto, Hideshiro [et al.]. In The Challenge of Information Technology (North Holland Publishing Co., 1983).

## APPENDIX

## JAPAN CATSS - SAMPLE SEARCH &amp; RECORD DISPLAY

アクセス・キー:

/ミマ, ユキオ

検索中

6 レコードが検索されました

ヒットリスト・コマンド:

\*6

|   |     |            |              |   |     |      |
|---|-----|------------|--------------|---|-----|------|
| 1 | DEM | 0099000104 | 聖女/三島//由紀夫   | n | 751 | 1951 |
| 2 | DEM | 0099000105 | 回想の三島由紀/日本学生 | n | 650 | 1971 |
| 3 | DEM | 0099000106 | 三島由紀夫/いいだ//  | n | 658 | 1970 |
| 4 | DEM | 0099000107 | 肉体の學校/三島//   | n | 751 | 1964 |
| 5 | DEM | 0099000108 | 三島由紀夫戯曲/三島// | n | 751 | 1962 |
| 6 | DEM | 0099000109 | 永すぎた春/三島//   | n | 751 | 1961 |

ヒットリスト・コマンド:

6

| RSN      | PTC  | OPN  | DFC     | DCR     | DCH     | TCH   | LNG |
|----------|------|------|---------|---------|---------|-------|-----|
| 99000109 | updt | DEMA | 88May27 | 88May27 | 88May27 | 17:19 | eng |

STATUS org fin

|     |        |     |      |     |   |     |   |     |   |
|-----|--------|-----|------|-----|---|-----|---|-----|---|
| 01: | 880527 | 03: | 1961 | 09: | h | 80: | a | 81: | n |
| 85: | n      |     |      |     |   |     |   |     |   |

|       |          |      |                                         |
|-------|----------|------|-----------------------------------------|
| #U040 | .....    | 0001 | \$aDEM\$beng                            |
| #U090 | 001..... | 0001 | \$aPL/833/I7N28                         |
| 251   | .....    | 0001 | \$a永すぎた春\$f三島由紀夫//著                     |
| 270   | .....    | 0001 | \$a[東京]\$b新潮社\$d[1961]                  |
| 275   | .....    | 0001 | \$a235 p.\$b15cm.                       |
| 280   | .....    | 0001 | \$a新潮文庫\$b1451                          |
| 551   | .....    | 0001 | \$aナガスギタ ハル\$xNagasugita haru\$b永すぎた春   |
| 751   | .....    | 0001 | \$aミシマ, ユキオ\$xMishima, Yukio \$b三島//由紀夫 |

他には?

# Japan CATSS

## Logon Screen

JAPAN CATSS

UTLAS INTERNATIONAL / HARUZEN

口座番号: 5801, MZNA: CAT  
パスワード: TKWD

日時 16:11 25 JUL 1986 システムID= 01267 ライン= 004-X25A1-00

| メールの種類       | 新 | 旧 | 作成済 | 送付済 |
|--------------|---|---|-----|-----|
| (N) 利用者用     | 0 | 0 | 0   | 0   |
| (A) 典拠変更通知   | 0 | 0 | 0   | 0   |
| (I) 相互貸借/複写用 | 0 | 0 | 0   | 0   |
| (O) 発注用      | 0 | 0 | 0   | 0   |

こんにちは。  
アクセス・キー:

## Logoff Screen

|                        | セッション                | 計    |
|------------------------|----------------------|------|
| 書誌                     |                      |      |
| 抽出                     | 0                    | 0    |
| リファイル                  | 0                    | 0    |
| オリジナル                  | 0                    | 0    |
| キャンセル                  | 14                   | 14   |
| ドロップ                   | 0                    | 0    |
| 典拠                     |                      |      |
| 抽出                     | 0                    | 0    |
| リファイル                  | 0                    | 0    |
| オリジナル                  | 0                    | 0    |
| キャンセル                  | 0                    | 0    |
| ドロップ                   | 0                    | 0    |
| 最初に付与されたRSN:           | 0                    | 0    |
| 最後に付与されたRSN:           | 0                    | 0    |
| 接続時間:                  | 1:12                 | 1:12 |
| さようなら 5801 , MZNA: CAT | 日時 16:09 25 JUL 1986 |      |

# Japan CATSS

## Precise Searching

Precise search keys retrieve records by various control numbers. Precise searches are entered after one of the following system prompts:

- Access Key: アクセス・キー:
- Enter browse command: ブラウズ・コマンド:
- Enter hitlist command: ヒットリスト・コマンド:

| Search key Prefix | Search Type                                   | Field Indexed | Example                                   |
|-------------------|-----------------------------------------------|---------------|-------------------------------------------|
| NONE              | RSN<br>(Utlas record control number)          | RSN           | Access key: 99000083<CR>                  |
| I/                | ISBN<br>(International Standard Book Numbers) | 010           | Access key: 4062007223<CR><br>(no hyphen) |
| J/                | Japan National Bibliography Number            | 001J          | Access key: 85059224<CR>                  |
| A/                | Accession Number                              | U035          | Access key: A/372643<CR>                  |
| K/                | NDL (National Diet Library) Card Number       | 906           | Access key: K/78B05539<CR>                |



# Japan CATSS

## ISBN Search

アクセス・キー:  
I/4062007223 <CR>

検索中

1 レコードが検索されました  
ヒットリスト・コマンド:  
1 <CR>

| RSN       | PTC  | OPN  | DFC     | DCR     | DCH     | TCH   | LNG |
|-----------|------|------|---------|---------|---------|-------|-----|
| 990000083 | orig | DEMA | 86Apr23 | 86Apr23 | 86Apr23 | 15:39 | eng |
| STATUS    | org  | fin  |         |         |         |       |     |

01: 860423 03: 1985 09: h 80: a 81: ■  
85: n

|       |          |      |                                  |  |  |  |  |
|-------|----------|------|----------------------------------|--|--|--|--|
| 001J  | .....    | 0001 | \$a85059224                      |  |  |  |  |
| 010   | .....    | 0001 | \$a4062007223                    |  |  |  |  |
| #U035 | .....    | 0001 | \$a85015605                      |  |  |  |  |
| #U040 | .....    | 0001 | \$aDEMS\$eng                     |  |  |  |  |
| #U090 | 001..... | 0001 | \$a913.6/Ta743k                  |  |  |  |  |
| 251   | .....    | 0001 | \$a海図\$f田久保英夫//著                 |  |  |  |  |
| 270   | .....    | 0001 | \$a東京\$b講談社\$d1985.6             |  |  |  |  |
| 275   | .....    | 0001 | \$a210p\$b20cm                   |  |  |  |  |
| 360   | .....    | 0001 | \$a1500円                         |  |  |  |  |
| 551   | .....    | 0001 | \$aカイズ\$xKaizu\$b海図              |  |  |  |  |
| 677   | .....    | 0001 | \$a913.6                         |  |  |  |  |
| 685   | .....    | 0001 | \$aKH589                         |  |  |  |  |
| 751   | .....    | 0001 | \$aタクボ, ヒデオ(1928生)               |  |  |  |  |
|       |          |      | \$xTakubo, Hideo(1928生)\$b田久保//英 |  |  |  |  |
|       |          |      | 夫(1928生)                         |  |  |  |  |
| 905   | .....    | 0001 | \$aKH89-569                      |  |  |  |  |

他には?

# Japan CATSS

## BROWSABLE SEARCHING

Similar to a catalogue, the online browsable index allows users to browse backward or forward to find the required heading. Searching can be done for *names, titles, series* or *subjects*.

Browsable search keys are entered after one of the following prompts:

- Access Key:                      アクセス・キー:
- Enter browse command: ブラウズ・コマンド:
- Enter hitlist command:    ヒットリスト・コマンド:

| Search key Prefix | Search Type                   | Tag Indexed             | Example                  |
|-------------------|-------------------------------|-------------------------|--------------------------|
| tag/              | Specific tag Search           | Any tag listed below    | アクセス・キー:<br>551/タイズ<CR>  |
| 5/                | Titles only<br>(incl. series) | 551-559, 580<br>590-599 | アクセス・キー:<br>5/タイズ<CR>    |
| 6/                | Subjects only                 | 650,658                 | アクセス・キー:<br>6/タイザ<CR>    |
| 7/                | Names only                    | 751-759, 791-799        | アクセス・キー:<br>7/ミヤ, ユキ<CR> |
| /                 | General search                | All tags listed above   | アクセス・キー:<br>/ミヤ, ユキ<CR>  |

# Japan CATSS

## Hit Lists

アクセス・キー: /ニホリキシ <CR>

検索中

9 レコードが検索されました

ヒットリスト・コマンド: +9 <CR>

|   |     |            |              |   |     |      |
|---|-----|------------|--------------|---|-----|------|
| 1 | MZN | 0099100005 | 講座日本歴史/歴史学研究 | n | 658 | 1984 |
| 2 | RUL | 0000400012 | 岩波講座日本/国史研究会 | n | 658 |      |
| 3 | JMC | 0010000499 | 日本歴史の再発見/網野善 | i | 658 | 1985 |
| 4 | JMC | 0010000500 | 歴史の陥穽/瀬野精一郎  | i | 658 | 1985 |
| 5 | JMC | 0010001777 | 霧の中の真実/田中起美恵 | i | 658 | 1984 |
| 6 | JMC | 0010003052 | 歴史への招待/歴史への招 | i | 658 | 1984 |
| 7 | JMC | 0010006870 | 要説史料日本史/要説史料 | i | 658 | 1984 |
| 8 | JMC | 0010008092 | 「日本史」異議あ/木屋隆 | i | 658 | 1985 |
| 9 | JMC | 0010010820 | 歴史の夜咄/司馬遼太郎  | i | 658 | 1985 |

ヒットリスト・コマンド: N

# Japan CATSS

## Citation Structure

|                                                                                                                       |     |            |       |        |   |   |     |    |      |
|-----------------------------------------------------------------------------------------------------------------------|-----|------------|-------|--------|---|---|-----|----|------|
| <div> <div>Tag where match found -</div> <div>ISBD - from 1085 -</div> <div>Encoding level - from 1084 -</div> </div> |     |            |       |        |   |   |     |    |      |
| 5                                                                                                                     | DEM | 0099000052 | 海図/   | 田久保//英 | 1 | i | 551 | CA | 1977 |
| Citation Number                                                                                                       | WHO | RSN        | TITLE | AUTHOR |   |   |     |    |      |
| <div> <div>Medium - from 1080 -</div> <div>Date of publication - from 1003 -</div> </div>                             |     |            |       |        |   |   |     |    |      |

### Medium 1080 values

If a or b, citation displays blank

If c or f, citation displays "CA"

a = language materials, printed  
b = language materials, manuscript  
c = cartographic materials, printed  
f = cartographic materials, manuscript

### Encoding Level 1084 values

✓ = full level

1 = sublevel 1

2 = sublevel 2

### ISBD 1085 values

✓ = full ISBD

n = non ISBD

i = Incomplete ISBD

# Japan CATSS

## Indexing Structure

### The Structure of Browsably Indexed Tags

|     |            |             |             |                   |            |
|-----|------------|-------------|-------------|-------------------|------------|
| 751 | .....      | 0001        | \$aミシマ, ユキオ | \$xHishima, Yukio | \$b三島//由紀夫 |
| Tag | Indicators | Occ.<br>No. | Kana        | Roman             | Kanji      |

### The Structure of the Browsable Index

|                |                  |
|----------------|------------------|
| 1.             | 2.               |
| ミシマユキオ<br>Kana | 三島//由紀夫<br>kanji |

# Japan CATSS

## Displaying Records - Matching Heading

アクセス・キー: /hix

検索中

3 レコードが検索されました

ヒットリスト・コマンド: \*3

|   |     |            |               |   |     |      |
|---|-----|------------|---------------|---|-----|------|
| 1 | SUL | 0001100008 | 海図/田久保英夫(1928 | n | 551 | 1985 |
| 2 | JMC | 0010010296 | 海図/田久保英夫      | i | 551 | 1985 |
| 3 | DEM | 0099000083 | 海図/田久保英夫(1928 | n | 551 | 1985 |

ヒットリスト・コマンド: l:disp:loc

|         |      |      |         |         |         |       |           |     |
|---------|------|------|---------|---------|---------|-------|-----------|-----|
| RSH     | PTC  | OPH  | DPC     | DCR     | DCH     | TCH   | SHR       | LNG |
| 1100008 | updt | SULA | 86Apr23 | 86Apr23 | 86Jun17 | 16:54 | 990000083 | eng |
| STATUS  | fin  |      |         |         |         |       |           |     |

01: 860423 03: 1985 09: h 80: a 81: ■

85: n

|       |          |      |                                                                  |
|-------|----------|------|------------------------------------------------------------------|
| 001J  | .....    | 0001 | \$a85059224                                                      |
| 010   | .....    | 0001 | \$a4062007223                                                    |
| 251   | .....    | 0001 | \$a海図\$f田久保英夫//若                                                 |
| 270   | .....    | 0001 | \$a東京\$b株式会社\$d1985.6                                            |
| 275   | .....    | 0001 | \$a210p\$b20cm                                                   |
| 360   | .....    | 2001 | \$b1500円                                                         |
| 551   | .....    | 0001 | \$aサイズ\$xKaizu\$b海図                                              |
| 677   | .....    | 0001 | \$a913.6                                                         |
| 685   | .....    | 0001 | \$aKH589                                                         |
| 751   | .....    | 0001 | \$aタクボ、ヒデオ(1928生)\$xTakubo,<br>Hideo(1928生)\$b田久保//英夫(<br>1928生) |
| 905   | .....    | 0001 | \$aKH89-569                                                      |
| *U035 | .....    | 2001 | \$a85015605                                                      |
| *U040 | .....    | 2001 | \$aSUL\$beng                                                     |
| *U090 | 001..... | 2001 | \$a913.6/T743h                                                   |

他には?



# Japan CATSS

## Browsing from Closest Heading

アクセス・キー: /済 <CR>

検索中

0 史記を中国語で読む

ブラウズ・コマンド: +12 <CR>

1 子規記念博物館松山市立

2 磁気記録

3 磁気記録最新技術と装置機器

4 磁気圏

5 色彩工学コンファレンス論文集

6 色彩と幻想

7 子規山脈の人々

8 志木市

9 志木市史

10 敷地天神山遺跡群

11 式子内親王1201

12 式子内親王集

ブラウズ・コマンド: -1 <CR>

11 式子内親王1201

ブラウズ・コマンド: D <CR>

検索中

1 レコードが検索されました

ヒットリスト・コマンド: +1 <CR>

1 CUN 0000100014 式子内親王集/式子内親王( n 751 1959

ヒットリスト・コマンド: 1: <CR>

RSN PTC OPN DFC DCR DCH TCH LNG

100014 orig CUNA 86Jun25 86Jun25 86Jun25 15:39 eng

STATUS org fin

01: 860625 03: 1959 09: h 80: a 81: n

85: n

251 ..... 0001 \$a式子内親王集, 守覚法親王集\$f国島章江  
//校 解説

270 ..... 0001 \$a東京\$b古典文庫\$d1959

275 ..... 0001 \$a286p

280 ..... 0001 \$a古典文庫\$b144

350 ..... 0001 \$a注: 本書は宮内庁書陵部蔵の写本 =斎院  
御集(式子内親王集)及び...

551 ..... 0001 \$aシキシ ナイシンノウ シュウ\$xShikishi  
naishinn^o sh^u\$b式子内親王集

551 ..... 0002 \$aシュカクホウ ナイシンノウ シュウ\$xShukakuh^o  
naishinn^o sh^u\$b守覚法内親王集

751 ..... 0001 \$aシキシ ナイシンノウ(?-1201)\$xShikishi  
naishinn^o(?-1201)\$b式子内親王(?-1201)

751 ..... 0002 \$aクニシマ, アキエ\$xKunishima, Akie  
\$b国島//章江

続けますか? (Y=はい, N=いいえ)

.LOC

=U035 ..... 0001 \$a168423

=U040 ..... 0001 \$aCUN\$beng

=U090 001..... 0001 \$a918/4931\$d#144

他には?

# Japan CATSS

## Browsing to Desired Heading

アクセス・キー: /カタ /

検索中

0 長田エミ

ブラウズ・コマンド: +5

1 中田平

2 中田耕治

3 中田小学校遺跡

4 中田信哉

5 永田武

ブラウズ・コマンド: D

検索中

1 レコードが検索されました

ヒットリスト・コマンド: +1

1 JMC 0010001371 Proceedings of/<永田武> i 751 1984

ヒットリスト・コマンド: /+1

1 中谷和夫

ブラウズ・コマンド: +2

2 中谷一正

3 中谷順一

ブラウズ・コマンド: D

検索中

1. レコードが検索されました

ヒットリスト・コマンド: +1

1 JMC 0010010828 南帝由来考/中谷順一 i 751 1985

ヒットリスト・コマンド: BROW

ブラウズ・コマンド: +2

1 中谷真弓

2 仲田紀夫

ブラウズ・コマンド: N

アクセス・キー:

Return to heading level

# Japan CATSS

## Truncation

アクセス・キー: /件\*リスレキ  
検索中

1 レコードが検索されました  
ヒットリスト・コマンド: +1  
1 JMC 0010003082 イギリス国民の歴/和田勇 i 658 1985  
ヒットリスト・コマンド: 1;.658  
658 .0..... 0001 \$aイギリス//レキシ\$xIgirisu//Rekisi\$bイ  
ギリス//歴史

他には?

CANC

B10-003-082 キャンセルされました 25-JLT-86 Logon 5801,MZNA:CAT

アクセス・キー: /件\*リスレキ

検索中

2 レコードが検索されました  
ヒットリスト・コマンド: +2  
1 JMC 0010003082 イギリス国民の歴/和田勇 i 658 1985  
2 JMC 0010003083 イギリス近代史研/小林照 i 658 1985  
ヒットリスト・コマンド: 2;.658  
658 .0..... 0001 \$aイギリス//レキシ//キンダイ  
\$xIgirisu//Rekisi//Kindai\$bイギリス//歴史  
//近代

他には?

## **Japan CATSS**

### **Filing of Alphabets**

1. Digits
2. Roman Alphabet
3. Greek
4. Cyrillic
5. Japanese

Utlas International Canada  
JAPAN CATSS

## Kanji Specific Search

### Regular Search

ヒットリスト・コマンド: /セイザイ  
検索中  
2 レコードが検索されました  
ヒットリスト・コマンド: +2  
1 JMC 0010001545 石川県における木材需/石 i 658 1985  
2 JMC 0010011182 製剤の安定化とそ/日本薬 i 658 1981  
ヒットリスト・コマンド: 1  
RSN PTC OPN DFC DCR DCH TCH LNG  
10001545 orig JMC1 43Jun14 85Oct04 86Jun27 13:03 eng  
STATUS org  
他には?  
<658  
658 .0..... 0001 \$aセイザイ  
\$xSeizai  
\$b製材

他には?  
canc:rsn:2  
B10-001-545 キャンセルされました 04-JLT-86 Logon 5801,MZNA:CAT  
2 レコードが検索されました  
RSN PTC OPN DFC DCR DCH TCH LNG  
10011182 orig JMC1 43Jun14 85Nov25 86Jul02 17:16 eng  
STATUS org  
他には?  
<658  
658 .0..... 0001 \$aセイザイ  
\$xSeizai  
\$b製剤

他には?  
canc  
B10-011-182 キャンセルされました 04-JLT-86 Logon 5801,MZNA:CAT

### Kanji Specific Search

アクセス・キー: /\$aセイザイ\$b製材  
検索中  
1 レコードが検索されました  
ヒットリスト・コマンド: +1  
1 JMC 0010001545 石川県における木材需/石 i 658 1985

Note: The kanji specific search reduces the number of hits from 2 to 1.



1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

68-MCULTP-8-E  
INF

THE CJK ENHANCEMENTS TO THE RLIN SYSTEM:

A REVIEW OF BASIC ISSUES

by

John W. Haeger

Vice President for Programs and Planning

The Research Libraries Group, Inc.

Stanford, California, USA

KEYWORD (S):

CJK Enhancements  
RLIN System  
Issues

For internal use only:

To be brought to meeting no:

Friday 22 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

159

207



米国の研究図書館グループ(RLG: Research Libraries Group)の図書館ネットワークシステム(RLIN: Research Libraries Network)における中国語、日本語、韓国語書誌情報への機能拡張にあたって直面した問題と解決に関する小史について述べる。

東アジア言語による書誌情報の入力、蓄積、伝送、検索、画面表示、印字出力といった機能を備えるこのプロジェクトは、1980年4月に開発を始めた。1983年9月に最初の“CJK”(中国語、日本語、韓国語)の書誌レコードがRLINシステムに入力された。1986年4月現在、北米の東アジア資料の約7割を所蔵する22の図書館が参加しており、約16万件的CJKレコードが蓄積され、69台のCJK端末が接続されて稼働している。地域的には米国、カナダを中心とし、英国のBritish Libraryとは特別の通信回線を通して2台のCJK端末が接続されている。本プロジェクトの最終段階としてCJKシソーラスの開発を進めており、今年の6月に完成する予定である。このシソーラスは、個々の漢字の属性についてのオンラインによる参照案内、RLINシステムが保有する漢字セットへの新しい漢字の追加、日本、中国、台湾、韓国の標準コード間のコード変換を実現するものである。

CJKプロジェクトでは、いくつかの経済上および設計上の基本的な決定事項があった。第1に新しいシステムを作るのではなく、既存のIRシステムを拡張するという決定である。そして中国語、日本語、韓国語が、これらの言語とローマ字の間で簡単に切り替えられて入力できるという決定である。このため、RLGはTranstech社のSintermと呼ばれる、漢字の構成要素(偏や旁など)を組み合わせた形で入力するシステムを採用した。重要なことは情報の検索であり、単にカード目録の模写ではなく、これを拡張して利用者がローマ字や漢字を自由に選択して検索できることであった。

プロジェクトで最も時間を要したのは漢字コードシステムの開発であった。問題は3種類の複雑な漢字の字体や字形であった。これらは相互に関連するものの異なったものであり、悪いことに、これらが安定していないことである。RLGはともかく新しい文字の追加はあってもREACC(East Asian Character Code)を凍結した。文字の変更などはシソーラスの変換テーブルで保守していく考えである。REACCはLCによって米国のNISO(National Information Standards Organization)に、さらにISOにも登録される筈である。REACCは書誌的なCJKのデータの国際的な情報交換における唯一のコード・システムであるといえる。

現在の端末機器の構成は、1981年に採用を決定したもので、機能的にも現在のものよりかなり制限されたものであった。プリンタもディスプレイも16×18より多くのドットでのCJK文字表示はできない。機種を選んだ1981年当時はまだ端末機が幼稚で、ミニ・コンピュータも安くはなかったのである。RLGではもっと安く機能的に秀れたもので、しかもRLINシステムで他の非ローマ字系言語(例えばCyrillic, Arabic, Hebrew など)も扱える端末機に改善する予定である。

最後に、RLGではこのシステムで3×5インチのいわゆる印刷カードの作成をサポートしない方針であった。その理由の1つは、当時印刷カードを作成できるようにプログラムを拡張することが非常に高価であったこと。もうひとつはRLGのスタッフや委員会のメンバーが、RLINは東アジア図書館に利用者が直接オンラインで目録をアクセスする方式(on-line public access)を実施させるシステムであり、カード目録はなくすものという信念をもっていたからである。しかし5年経過した現在、カード目録をクローズした館はわずか1館にすぎない。コロムビア大学をはじめ他の5、6館ではRLINシステムからのローマ字形のカードを注文し、これに手で漢字を書き入れているのである。このようなCJKカードへの参加館の継続的な要求は無視することができない。その解決方法は明確でないが、RLGとしてはその技術的解決を再評価しつつあるのが現状である。

RLINはネットワークを基盤とした中国語、日本語、韓国語の図書館資料を処理し、検索し得る、現在フル稼働している世界で唯一のシステムである。今後さらに利用者の要望を考慮し、より広範囲な書誌的環境を反映すべくシステム改善を進めていかねばならない。

# ABSTRACT

In April 1980, the Research Libraries Group began work to create Chinese, Japanese and Korean enhancements to its on-line bibliographic utility, the Research Libraries Information Network (RLIN). Economic and design decisions underpinning RLIN CJK have focused on (1) integration of a system in which information can be retrieved irrespective of the language or country of its publication; (2) the creation of an input system that could work easily with Chinese, Japanese and Korean and that could switch among the three and the roman alphabet; (3) better information retrieval to expand searching capabilities and to decrease the need for card catalogs; (4) the choice of a standard code for vernacular characters and the subsequent creation of REACC, and (5) hardware choices to reduce costs and to increase functionality.

Between April 1980 and September 1983, the Research Libraries Group, Inc. (RLG) developed enhancements to the Research Libraries Information Network (RLIN), its wholly-owned bibliographic utility, capable of supporting input, storage, transmission, search, display and printing of bibliographic records containing East Asian vernacular script. These enhancements have been amply described in papers by my colleagues Alan Tucker and Karen Smith-Yoshimura, and in articles which have been published in the library press. On September 12, 1983, the Library of Congress in Washington entered the first "CJK" record into the RLIN system, and shortly thereafter records were created by three other pilot libraries. As of April 1986, 22 institutions, collectively representing approximately 70% of the East Asian holdings in North America, use RLIN for most or all of their CJK vernacular processing, and the RLIN database now contains approximately 160,000 CJK records. Sixty-nine CJK terminals have been installed in the United States, Canada and the United Kingdom.

The final module of the on-line RLIN CJK system, the so-called CJK Thesaurus, is still under construction. The Thesaurus is an on-line reference guide to the attributes of individual Chinese characters, a resource from which new characters can be added to the active RLIN repertory, and a conversion table to and among the "national" standard codes for individual characters developed in Japan, China, Taiwan and Korea. Installation of the Thesaurus is scheduled for June 1986.

This preconference may be a good opportunity to revisit some of the basic economic and design decisions underpinning RLIN CJK. Almost seven years have passed, after all, since RLG committed itself to build support for East Asian scripts in RLIN, six years since the project was begun, and three since implementation. Design options and technology costs have changed, and enormous progress has been made in related projects worldwide.

1. Integration. RLG was committed from the outset to enhance a single, preexisting integrated library processing and information retrieval system, not to design a new, separate, or stand-alone system for North American collections of East Asian materials. All functions of RLIN's on-line system are supported for CJK, including all formats and all modules except Authorities. Philosophically, we believe this premise remains valid. We will not have a universal science of man and his works, drawn from all cultures, if technology mimics all the barriers which time, history, language and convenience have erected in the organization and classification of information. We need an infrastructure in which information on urban planning, for example, is retrieved irrespective of the language or country of its publication. As I wrote for the First Asian Pacific Conference on Library Science in 1983, "It makes no difference that most RLIN users do not read Chinese, Japanese or Korean, or that most North American students of East Asia will be principally interested in East Asian vernacular material. It is still true that silence is often mistaken for negative proof, that errors are committed because a lack of information is

assumed to mean that the information does not exist, and that it can often be very important to know that a body of literature on a particular subject exists in a language which one does not, for the moment, read."

There have been two inconvenient consequences of our insistence on integration, however. First, in order to ensure that every bibliographic record in RLIN could be displayed, at least in part, on every RLIN terminal, we developed the concept of core and extended fields, and established a requirement that data in core fields be entered in parallel pairs, one vernacular and one romanized. In this way, at least the romanized portion of every record (the common denominator of each record in the North American context) could always be displayed, no matter what the capabilities of the searching terminal. This requirement is costly (though probably not as expensive as some users were at first inclined to believe), and the putative cost has led to criticism of the requirement.

Second, RLG has been unable to export its "CJK system." CJK, as we have designed it, is inseparable from RLIN itself.

Furthermore, RLIN in its turn is temporarily inseparable from Stanford University. The underlying operating system software, notably the time-sharing and teleprocessing monitors, developed by and for Stanford, is not documented or supported in ways which make replication or export very practical. In order to utilize CJK (and therefore RLIN) on the RLG mainframe beyond the geographical confines of North America involves expensive telecommunications charges and service window constraints. Since



1983, therefore, RLG has had to respond negatively to inquiries from libraries around the world which have expressed interest in using the "CJK system." We have had to say that, for all practical purposes, it was not available outside North America. This year, the British Library in London has installed two CJK terminals on an experimental basis. This is possible only because a dedicated telecommunications line was installed for another purpose and because the BL, provisionally, was prepared to accept afternoon-only service. Late in 1986 or early in 1987, however, RLG expects to move off of the Stanford-developed operating software onto a vendor-supported infrastructure. If this transition is successfully accomplished, it will simultaneously make possible: (1) the export of the RLIN system, including CJK, to other processors in other parts of the world; and (2) the extension of the RLIN service day to nearly 24 hours. For those who remain interested in using RLIN for CJK processing, this will open new opportunities.

2. Input. From the outset, we attached very limited importance to the choice of an input system for vernacular characters. This may seem paradoxical, since the absence of an obvious character input device had been (and perhaps remains) the single largest obstacle to the development and acceptance of word processing and data manipulation systems in East Asia. However, the hard fact is that, despite extravagant claims, no system available in 1980 looked "ideal." RLG had only one immutable requirement: we needed a system that worked more or less as well for Chinese as for Japanese and/or Korean, and that could switch easily among



the three languages and the roman alphabet. Data entry speed seemed relatively unimportant: the creation of a bibliographic record has never been an exercise in speed-typing! In 1981, we chose a character-component entry system, Sinoterm, marketed by Transtech International Corporation of Natick, Massachusetts. While our users seem generally pleased with the Sinoterm system, and we may continue to use it, uniquely or in combination with others, for the foreseeable future, all other features of RLIN CJK are fully divisible from the input system. In fact, cognizance of the input system is limited to the CJK terminal. The RLIN applications and data base ignore it entirely.

3. Information Retrieval. By contrast, we attached great importance to information retrieval. We believed that RLIN should not merely replicate the capabilities of a card catalog, but expand them. In particular, we wished to permit searching by the user's choice of vernacular characters or romanization, and to permit the confinement of a search to a single choice among variant forms of a character, or conversely, its extension across all variant forms. With these capabilities, a user knowing the vernacular characters of an author's name, for example, would not need to sort through homophonous names along the way; a user knowing only the simplified or joyo kanji form of a character could still retrieve records embedding the same character in their more complex, traditional form or forms -- all without resort to any intermediate reference tools. RLIN CJK relies on special, additional indexes and a relational, multiplanar character coding scheme for these capabilities.

4. Standard Codes for Vernacular Characters. In 1980 RLG had no intention of establishing a new code for the machine representation of vernacular characters. In a paper written for the Canberra (Australia) Conference in 1981, I described the RLIN East Asian Character Code (REACC) as "a database management oriented approach to the East Asian character set problem" which "acknowledges the coexistence of overlapping 'standard' sets and allows translation between them." I emphasized that it was "not a new standard in its own right." Five years later, I am much less sure. No aspect of RLIN CJK development has, in the end, been more exasperating or time-consuming than the care and feeding of REACC. In part, we simply underestimated the magnitude of the task. Working across three complex orthographic environments, related but divergent -- and sometimes unexpectedly divergent on close examination -- has been excruciating. Worse, the base standards from which we worked -- Japanese Industrial Standard (JIS), the Chinese Character Code for Information Interchange (CCCII), Guojia Biaojun (GB2312-80), and the Korean Information Processing System (KIPS) -- have proven notoriously unstable. In some instances, the instability is a consequence of the early release of unfinished and uncorrected drafts. In other cases, changes made in the standard codes and duly promulgated, and completely reasonable in most user environments, turn out prejudicial to bibliographic data. With the latter, after all, a change in the prescribed graphic representation of a character introduced into an on-line system after a record has been created may result in an ex-post-facto divergence between the automated record and the original title page. When the number of CJK

records in the RLIN database was small, laundering the database to substitute new codes for old was conceivable. This is less and less true. REACC cannot be perpetually reconformed to its "parent" standards. Although new characters will be added when and as necessary, RLG now has decided to freeze REACC in all other respects. Copies of the code will be distributed soon by the Library of Congress' MARC Distribution Service, and LC will undertake to register REACC first with the U.S. National Information Standards Organization (NISO) and then with ISO. Once REACC is frozen with respect to changes, we will still attempt to keep the Thesaurus-based conversion tables up to date. Nevertheless, unless the agencies responsible for the "parent" codes issue their modifications in the form of explicit change lists, even this will become impossible. Neither RLG nor any other code user has the time to scan new versions of a printed codebook to locate and verify embedded changes. Absent some dramatic change soon in the behavior of standards agencies in East Asia, we face the real prospect that REACC itself may become the only viable international communications code for CJK bibliographic data. As my colleague Karen Smith-Yoshimura wrote in a paper for the Second Asian Pacific Conference on Library Science held last year in Seoul, "If East Asian MARC records are to be successfully loaded into RLIN one day, with the vernacular data correctly converted into REACC, then it is vital that . . . code mapping . . . [be] comprehensive and up-to-date. Similarly, if RLIN CJK records are to be distributed to East Asian

libraries, then it is equally essential to know which subset of the character set can be interpreted by the recipient."

5. Hardware. The current generation of RLIN CJK terminals is a relatively complex integration of LSI 11/23 minicomputers, Motorola monitors, special keyboards, and General Electric printers. The monitors and printers cannot manage a dot-matrix for CJK characters larger than 16x18. (For simplicity, the same matrix was used both for on-line display and for casual printing.) It is important to remember that, when this configuration was chosen in 1981, the IBM-PC was still an infant product and minicomputers were not cheap. The choices in 1981 were both more limited and less obvious than they would be now. Sometime in the next two years, RLG intends to reimplement CJK on a new terminal hardware base, most likely some version of the IBM-PC. This reimplementation will probably cut the cost of RLIN CJK hardware while simultaneously providing enhanced functionality. It will also place CJK in the same hardware environment with other non-Roman enhancements to RLIN -- those for Cyrillic, Arabic and Hebrew. It is important to emphasize that the principal motivation here is not economic. Even a 30% decrease in the hardware cost for RLIN CJK may not have much impact on bottom-line economics. Over a five year period, assuming full amortization, only about 16% of the total cost of CJK to an average user is hardware. This would mean that a 30% decrease in hardware costs would net to less than a 4% savings

against total operating costs. Savings to very low-volume users would be more substantial, however.

6. Printed Cards. Perhaps the most controversial of RLIN CJK's early design decisions was non-support for printed 3x5 cards. We took this decision for two reasons: (1) the costs of enhancing RLIN's general purpose card production program to generate vernacular CJK cards looked extremely high in 1980. Card production could not have been driven by the 16X18 dot matrix graphics used for on-line display and casual printing, so entirely new font work would have been required. Good vendor-supported tools for digitization were not available. (2) RLG staff, Board and at least some members of its East Asian Program Committee believed that RLIN itself could be used as an on-line public access catalog for East Asian libraries, eliminating the need for cards. While such use of RLIN was unthinkable in general research libraries both because of system load and terminal access requirements, neither constraint appeared valid for East Asian libraries in the U.S., where patron traffic is generally light. Five years later, only the Hoover Institution at Stanford has actually closed its card catalog. Columbia University and several others order from RLIN romanized cards on which they handwrite vernacular characters! Several shortcomings of the RLIN system (e.g., inability to confine a search to a library's own records and the complexity of the Sinoterm keyboard) have been cited as impediments to the use of RLIN as an OPAC-substitute. Be that as it may, the demand for CJK cards



continues unabated. The solution is not obvious, and RLG is presently reassessing its technical alternatives.

As this paper is written, RLIN is the only on-line, network-based, library processing and information retrieval system capable of handling Chinese, Japanese and Korean vernacular which is fully operational anywhere. It is likely that, by the time the paper is presented in August, that situation will have fundamentally changed. In the U.S., OCLC is scheduled to be in pilot operation this summer. Developments at UTLAS, led by their interest in the Japanese market, continue. The system developed by Taipei's National Central Library may move soon to a more flexible hardware environment. Many other developments may be reported to this preconference. These will fundamentally alter the landscape. The features of various systems will be compared. A multiplicity of user requirements, reflecting a wide range of bibliographic environments, is likely to become better defined. Data created on separate and different systems will need to be communicated and exchanged, and this will create its own array of potentially nasty problems. Whatever happens, RLG is grateful to the Andrew W. Mellon and Ford Foundations and to the National Endowment for the Humanities, to its members and consultants, and to its many friends in Asia -- many of them here in Japan -- for having made our early work possible.

April 1986  
Stanford, California





1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

69-MCULTP-9-E  
INF

Bibliographic Control and Library  
Automation in Japan

By

Shôjirô Maruyama

Library Counsellor

National Diet Library

Tokyo, Japan

KEYWORD(S):

Japanese bibliographic data

Cataloging rules : Japan

JAPAN/MARC

Library automation : Japan

For internal use only:

To be brought to meeting no:

Friday 22 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

日本における書誌コントロールと図書館機械化につき、まず歴史的に書誌コントロールの発展を紹介し、次に日本の目録法の特徴をいくつか具体的にあげる。また、日本における書誌情報に関連した図書館機械化の状況にふれ、書誌情報ユティリティの現状と将来性について言及する。

書誌コントロールについては、戦前の状況を若干示すとともに、1948年の国立国会図書館(NDL)の設立と同年から開始された全国書誌の刊行、NDLおよび学術雑誌の全国総合目録編纂事業を最初にあげる。

欧米諸国で普及している、出版界との協力による図書館中への書誌情報すり込み計画(CIP)が日本で実現されない理由を考察、印刷カード・サービス、民間業者による購入図書への目録カード添付サービスにも言及する。

最近多くの図書館が活用しだした、取次業界が作成頒布している民間MARCテープやオンライン出版情報サービスについても紹介をする。

次に日本語資料の目録法における特徴をいくつか説明する。NCR(日本目録規則)の歴史にふれ、基本記入制の導入と現行の予備版における記述ユニット制を説明する。特に日本語標目における片かな(もしくはローマ字)と漢字の二重性につき、具体的に説明し、1955年以来提唱されてきた記述独立方式の経緯についても紹介する。和図書の出版慣行が洋書と相違していることを指摘し、主要な目録情報源の標題紙以外(奥付等)への分散、それに対応した日本独特の目録慣行についても簡単にふれる。

国際的な目録法原則(標目に関するいわゆるバリ原則)やISBD(国際標準書誌記述)とNCRの整合性、和図書や日本語の特性に基づく独自性についても言及し、特にISBD区切り記号の取扱い、NCR新版予備版は複数記入制の記述ユニット記入を前提としているので、予備版における「単一記入制目録のための標目選定表」にふれている。

日本における図書館機械化では、特に漢字のコンピュータ処理を重点的にとりあげた。1970年代初頭からのNDLの図書館機械化を紹介し、漢字のJIS規格制定前に決定され、独自使用されたNDL専用の漢字コード表“NDL-70コード”、JAPAN/MARCで使用されているJIS規格“JIS C6226-1978”との関連およびその外字についてふれている。(1986年までに追加された外字は700字以上)

次に、UNIMARCフォーマットとJAPAN/MARCフォーマットの比較を行い、実例を用いて漢字記述データと片かなやローマ字形の標目データのリンク用タグづけについて説明する。

JAPAN/MARCテープの収録範囲やレコード数、1986年現在のテープ頒布サービス予約者の内訳を示し、1983年に日本図書館協会が行なった公共図書館と大学図書館に対する機械化の調査報告の結果を引用し、日本の図書館はコストは高いが漢字処理システムの導入を選ぶ傾向があることを指摘する。

標準化の普及については、ISBNやISSN付与率、NDC(日本十進分類法)が全館種の4万の図書館の96%で使用されていることなどを紹介する。

最後に、日本における書誌情報ユティリティの状況と方向について、JAPAN/MARCオンライン・ネットワークのテスト的な運営、UTLAS International Canada社のプロトタイプ、学術情報センターの動きにふれ、日本の図書館界がオンライン総合目録、書誌情報の共有によって新しい時代をむかえることを指摘。

最後に、漢字圏の共通フォーマットとして、UNIMARC漢字版、共通漢字コード表、共通目録法原則の研究、特に分類記号、件名標目等の主題関連データの互換性確立のための調査などの必要性をあげる。

Bibliographic Control and Library Automation in Japan,  
by Shôjirô Maruyama (National Diet Library)

Abstract: A historical background and the present state of bibliographic control in Japan is introduced. The establishment of the National Diet Library (NDL) and its bibliographic products, such as a national bibliography and JAPAN/MARC (J/MARC), etc. are described. Afterwards, Japanese standard cataloging rules, NCR, features of Japanese cataloging rules and J/MARC format are explained. Then, the following topics are enumerated: 1) Problems and solutions of Kanji data processing, 2) Library automation in Japan, 3) Proliferation of Private sector MARCs, 4) Standardization, 5) Prospect of bibliographic utilities.

## 1. Bibliographic control in Japan

### 1. 1 National bibliography and the National Diet Library

Under the document, "Guidelines for the national bibliographic agency and the national bibliography", prepared by the IFLA International Office for UBC in 1979, the national bibliography can be defined as the cumulation of the authoritative and comprehensive records of the national imprint of a country, published regularly, and with the least possible delay.

The National Diet Library (NDL), the national library of Japan created in 1948, is the culmination of the Imperial Library and the attached libraries of both Houses of the National Diet. The NDL is responsible for publishing a catalog or index of the publications issued within Japan during the year, at periodic intervals of not more than one year. From 1948 onwards, the library started issuing a monthly national bibliography. In 1955, the title and frequency of this publication was changed to "Japaneses national bibliography weekly list". Based upon these publications and supplemented by further investigation of missing titles, annual edition of "Japanese national bibliography" has been compiled and published since 1951.

### 1. 2 A glimpse of bibliographic control before 1948

Under the Publication Act in 1867, the Government assumed control over all publications. All publishers were obliged to submit their newly published books to the Ministry of Home Affairs for censorship review. Consequently, the records of that time were fairly complete. Both the Government and some publishers produced monthly or annual catalogs of new publications. But, these were only published intermittently. The Imperial Library started publishing "Report of the Imperial Library", at almost monthly intervals, in 1908. This report, actually a classified catalog of books newly acquired by the

Library was the most comprehensive national bibliography at that time, until it ceased to appear in 1941, at the outbreak of the War.

### 1. 3 Union catalogs in Japan

After creation of the NDL, the Library started to compile and maintain the national union catalogs of books, and with cooperation of other libraries, "The union catalog of foreign books" started publication on yearly basis, in 1954. In 1953, the Ministry of Education had begun compiling and publishing a "Union list of scientific periodicals". This list consists of 4 sections shown as follows.

- Humanities and social sciences in Japanese language
- Humanities and social sciences in European languages
- Natural sciences in Japanese language
- Science and technology in European languages

At present, the bibliographic information of this union list has been converted into machine readable form and is maintained as a data base by the National Organization for Science Information System (SIS), former Center for Bibliographic Information (CBI) of the University of Tokyo.

### 1. 4 Unattainable CIP program

The CIP (Cataloging in Publication) program has been unsuccessful in Japan. To explain this situation, I'd like to make a few comparisons.

- 1) The schedule of Japanese publishers is tighter compared with those of American and European publishers.
- 2) The problem of implementing this program due to the library's inflexible working procedure in manual system.

#### 1. 5 Printed catalog card services

In 1949, the NDL began preparatory work to distribute printed catalog cards for general public. But, the actual distribution service didn't start until December 1950. Since April 1980, the production of the catalog has become more efficient by the use of an automated system based on the J/MARC data base. The Japan Library Association (JLA) also started its own printed catalog card distribution service in August 1952 and in 1957, the JLA's new cards-with-basic books program has started. Some libraries have adopted this labor saving device as their option since then. Dating the 1970's, the public library service expanded its operation in Japan. In 1972, the Osakaya Co. offered the first private sector cards-with-books service to a public library in Tokyo. Now, JLA's printed catalog cards are also produced using a private sector MARC data base.

#### 1. 6 MARCs in Japan

The NDL has created and distributed the JAPAN/MARC (J/MARC) tapes containing machine readable national bibliographic information since April 1981. In June 1979, Nihon Shuppan Hanbai, Inc. started distributing the so-called Nippan MARC as a prototype. Thereafter major book vendors of Japan began producing and distributing private sector MARC tapes which adding approximately 25,000 records per year. Those MARC records are created from the books which those vendors are handling. The private sector MARC data bases are accessible to bookstores using online network facilities. In 1986, new online systems utilizing new publications were also developed. The main features of these systems are as follows.

- 1) Joint venture of big book vendors, book jobbers, and index publishing companies.
- 2) Records containing summaries of the books.



## 2. Features of Japanese language materials cataloging

### 2.1 History of the Nippon Cataloging Rules (NCR)

The first edition of the NCR was compiled by the Cataloging Committee of the League of Young Librarians, and was published in 1942. The framework of this Rules was influenced by first Anglo-American cataloging rules of 1908. After the establishment of the NDL, the 1952 edition of the NCR was published as the national standard for cataloging and the NDL printed catalog cards were then produced according to these rules. The NCR 1965 edition was revised according to the Paris Principles, and until this edition, the NCR was based on the principle of main entry. A newly revised edition of the NCR based on ISBDs published in 1977, introduced a unique concept named "Description Unit System" (DUS, a literal translation of the Japanese terms). Purposes for introducing this system are:

- 1) Since the bibliographic description unit based on the ISBD is able to uniquely identify each item. The unit may stand alone as a record which can be used as the basis of a catalog entry.
- 2) The bibliographic record of Japanese books has to contain both Kanji (Chinese character) and Kana (phonetic Japanese alphabets) or Romanized data for identification and filing purposes respectively. In Western language catalog entries, a heading in the Roman alphabet works well for both identifying the author and showing the filing position. But in our Japanese headings, we have to use Kana or Romanized heading and Kanji heading together. Because homophones are very numerous in Japanese names and sole Kana or Romanized form names without Kanji form would not be sufficient to identify an author.

### 2.2 New structure of the NCR

In our new NCR, rules for description precede rules

headings. This structure coincides with Description Unit System (DUS). The bibliographic record according to the DUS consists of a description unit and following heading indication unit. In our new NCR, we have not used the term "tracing", instead we have used the term "heading indication unit". Because, we have adopted "alternative heading entry system as mentioned in the AACR2, 0.5. The description unit of DUS corresponds to the Descriptive and Notes blocks of the UNIMARC, and the heading indication unit corresponds to Related title, Subject analysis, and Intellectual responsibility blocks, respectively.

We have also added new rules to this new NCR, namely "levels of detail in the description" and "bibliographic levels" in preparation for the coming online cataloging era.

## 2. 3 Forerunners of the Description Unit System

In 1955, despite main entry principle, a group of catalogers started advocating the complete separation of the descriptive part of the catalog entry from its heading, because about 60 percent of Japanese public libraries use Kana headings, and over 80 percent of university libraries in Japan use Romanized headings, it was considered expedient to use unit cards without main entry headings.

The DUS system of the new NCR is intended to fit the multiple entry cataloging. It should be understood that the DUS is not based on the title main entry principle. Because, on past controversy of author vs. title main entry, the point at issue in dispute didn't stand on multiple entry cataloging.

## 2. 4 Conformities with and deviations from international standards

The main bibliographic information of Japanese books is not concentrated in the title page as in Western books but sometimes the greater part of information will be found in the colophon. The new NCR decided that the prime source of the

title and statement of responsibility area should be the title page (including the verso of the page), colophon, spine and cover. If the title varies in the above mentioned sources, the title proper should be in the form most often used in those four sources. When the title is given in different forms in these four sources, we should use the title on the title page, colophon, spine and cover in the prescribed order of preference.

The new NCR has been revised according to ISBD principles and framework of provision but, we have hesitated to some extent in adopting ISBD punctuation, because we restrict the use of punctuation to the minimum in Japanese sentences. So, use of punctuation had been also restricted at a minimum in our cataloging practice. Therefore, we have provided an alternative rule of allowing no use of ISBD punctuation for local purpose.

The new NCR's objective is with multiple entry cataloging, but sometimes we are required to make single-entry listing. The choice of heading according to the Paris Principles are considered very useful, in this case. We have provided a list called "Choice of the Main Entry for Single-Entry Listing - a Schematic Table" based on the Paris Principles.

### 3. Library automation in Japan

#### 3. 1 NDL and library automation

The NDL started its library automation project using its own computer system in 1971 with the immediate objective of processing Japanese bibliographic data. At that time, Kanji data processing equipment was very expensive. The Kanji processing component of its system costed nearly a half of total hardware expences. From January 1977, the NDL has

started working toward building up of a data base on monographs currently published in Japan. The initial step undertaken was the automated compilation of the "Japanese national bibliography weekly list". The first computerized issue was published in January 1978. This marked a departure from the conventional process of manual editing, eventually leading to the growth and improvement of a database, and finally to the development of the JAPAN/MARC (J/MARC) data base.

### 3. 2 NDL's Japanese character set

Prior to the establishment of the national standard, i. e. JIS code, the NDL developed its own character set and code (i. e. NDL-70 code). A committee working from 1969-1971, conducted an overall survey of Kanji use frequency in bibliographic records. It was agreed, first of all, that Kanji designated by competent authorities as "practical-use characters" should be chosen for inclusion. These totalled 1,968 characters at that time. In July 1970, a total of 4,567 Kanji was chosen to be included in the NDL-70 code. To these were added 1,150 non-Kanji characters comprised of Kana, Roman and Cyrillic alphabets, numerals, and other symbols, bringing the total figure to 5,717 characters. The massive bibliographic records so far accumulated by the NDL in its data base are based on the NDL-70 code. However, to put the data to outside use, the data must be converted to the standard JIS code. In the J/MARC tape, conversion is necessary from NDL-70 code to a lightly extended form JIS (Japan Industrial Standard) code.

In 1974, the Agency of Industrial Science and Technology, Ministry of International Trade and Industry took the initiative toward the establishment of a national standard for Kanji character code. The draft prepared by an ad hoc specialists committee was submitted to and approved in 1978, by the Agency

as JIS C6226-1978. (A revision followed in September 1983) The committee finally determined 6,349 characters as necessary. These characters are divided into two groups determined by their frequency of use in various Japanese texts. The first group of the code consists of 2,965 characters with higher frequencies and the second group consists of an additional 3,384 characters with lower frequencies. As is often the case, some characters in J/MARC record are not in the JIS code. Accordingly, we give such characters new individual code numbers whenever it is found appropriate. In the middle of 1986, a total of over 700 Kanji characters were given such new additional code numbers. Every year, we find it necessary to give about 40 Kanji characters new code numbers.

### 3. 3 J/MARC format

In its efforts to develop the J/MARC format, due consideration has been given to conformity with the UNIMARC format issued by the IFLA International Office for UBC. The J/MARC records are mixture of Kanji data and either Kana or Romanized data. This format has to be hospitable to both Kanji data using 2 bytes and Kana or the Romanized data using 1 byte. In addition, in the J/MARC format, to tie up Kanji data and corresponding access points in Kana or Romanized forms, a linking technique as shown table 1 (in p. 12) must be used.

### 3. 4 Coverage and distribution of J/MARC

At present the J/MARC data base contains about 600,000 records accumulated over the past nine years. The average current input per year is over 60,000 records, of which official publications, issued by central and local governments, take up around 20 percent.

The J/MARC tapes have been distributed on a subscription basis at weekly intervals. As of 1986, the total number of J/MARC subscribers has reached to 35 with a breakdown as follows: National universities (6), Private universities (12), National research museum (1), Public library (1), Computer service bureaus (5), Major printing companies (2), Bibliographic data producers (8).



At the moment, sophisticated Kanji processing systems are very expensive for small libraries. Therefore, some libraries are getting their printed catalog cards and printed catalogs from some contracted computer service bureaus or printing companies.

### 3. 5 Automation of public and university libraries

In 1983, the Japan Library Association conducted a nationwide survey on library automation in public and university libraries. Some remarkable findings from it are enumerated as follows:

- 1) 131 (15.7 percent) public libraries have already automated their operations.
- 2) 155 (48.6 percent) university libraries are using computers for information retrieval and in-house management.
- 3) Most libraries prefer sophisticated and rather expensive Kanji processing systems to simple and cheaper alphanumeric data processing systems.

### 3. 6 ISBN, ISSN and private sector MARC services

In Japan a system called the Japanese Book Code was in use from 1970 to 1979, and commercial publications with such code numbers made up approximately 70 percent of the total production, until the Japan Publishers Association decided to adopt ISBN in 1980. In 1986, over 80 percent of new commercial publications have been assigned ISBN. On serial publications, the NDL is a national center for the ISDS (International Serials Data System). It has been operational since 1976. Now, over 60 percent of Japanese serial publications are issued with ISSN (International Standard Serial Number).

Standardized cataloging tools, such as the NDC (Nippon Decimal Classification), the NCR, etc. are widely used all over Japan. For example, various editions of a standard classification scheme, namely the NDCs are used in 96 percent of our 40,000 libraries.



At least 200 libraries are now using the above-mentioned private sector MARC tapes, because of the updatedness of their informations and the MARC format of these tapes are convertible to J/MARC format.

#### 4. Present state and future prospects of bibliographic utilities

##### 4. 1 J/MARC online networks

A prerequisite to providing Kanji data service online is that a user be equipped with a Kanji terminals. Fortunately, the formulation of Kanji JIS code has been a big step forward, while an advance in technology has led to rapidly lowering the cost of equipment.

Experimental J/MARC online bibliographic information services are conducted by six institutions in Japan. The UTLAS Inc. of Canada also has been conducting experimental online service using the J/MARC data base. Now, over ten university libraries in Japan are users of the UTLAS service.

The National Center for Science Information System, new name of CBI, was established in 1984 to act as a bibliographic utility for college and university libraries in Japan. The CBI preceded the online bibliographic service, and under the Plan for the Science Information System of the Ministry of Education, its online system has been operational since 1979. Now 11 national and private universities have joined this system.

##### 4. 2 Online systems and bibliographic utilities

Automation of Japanese libraries is entering into a new era. The formation of cooperative library networks, the build-up of online union catalogs, and the sharing of bibliographic information and library resources through networks are pressing tasks ahead for us. In order to build up the MARC online network as a bibliographic utility, first priority has to be given to the development of a decentralized voluntary network.

When a national online and shared cataloging network gets started, data input from participating libraries will be able to enrich our J/MARC data base. Without cooperative support from other libraries, it seems impossible that the National Library will be able to provide complete bibliographic informations on the current domestic publications.

#### 4. 3 Kanji bibliographic information in international scene

Considered a very important task ahead of us is, with the close cooperation of countries concerned, to build up an international interchange network of Kanji bibliographic information. A need for Kanji bibliographic information exists in other countries outside the Far East. In addition to the RLIN-CJK Project, the OCLC and UTLAS have started their own CJK projects. To facilitate data conversion, development of a Kanji version of UNIMARC and a common Chinese character set is needed. The following topics are also considered urgent in promoting interchange of bibliographic data.

- 1) Standardization of bibliographic data, e. g. establishing common cataloging principles.
- 2) Study of data compatibility/convertibility issue, in case of classification systems and subject headings, etc.

Comparative table of UNIMARC vs.  
J/MARC tags

| UNIMARC                                        |                                                            | J/MARC                                             |                                 |                                 |
|------------------------------------------------|------------------------------------------------------------|----------------------------------------------------|---------------------------------|---------------------------------|
|                                                | English<br>Bibl. Data                                      | <u>Kanji</u><br>Bibl. Data                         | <u>Kana</u><br>Bibl. Data       | Romanized<br>Bibl. Data         |
| Descriptive<br>Information<br>Block            | Tag 200\$a:<br>Title proper                                | Tag 251\$a:<br>Title proper                        |                                 |                                 |
|                                                | Tag 200\$e:<br>Other title<br>information                  | Tag 251\$b:<br>Other title<br>information          |                                 |                                 |
|                                                | Tag 200\$f:<br>First state-<br>ment of res-<br>ponsibility | Tag 251\$f:<br>Statement of<br>responsi-<br>bility |                                 |                                 |
| Related<br>Titles<br>Block                     | (no room<br>for title<br>proper)                           |                                                    | Tag 551\$a:<br>Title<br>proper  | Tag 551\$x:<br>Title<br>proper  |
| Intellectu-<br>al Responsi-<br>bility<br>Block | Tag 700\$a:<br>Personal<br>name                            | Tag 751\$b:<br>Personal<br>name                    | Tag 751\$a:<br>Personal<br>name | Tag 751\$x:<br>Personal<br>name |
| Descriptive<br>Information<br>Block            | Tag 200\$c:<br>Second<br>title<br>proper                   | Tag 252\$a:<br>Second<br>title<br>proper           |                                 |                                 |
|                                                | Tag 200\$f:<br>First state-<br>ment of res-<br>ponsibility | Tag 252\$f:<br>Statement of<br>responsi-<br>bility |                                 |                                 |
| Related<br>Titles<br>Block                     | (no room<br>for title<br>proper)                           |                                                    | Tag 552\$a:<br>Title<br>proper  | Tag 552\$x:<br>Title<br>proper  |
| Intellectu-<br>al Responsi-<br>bility<br>Block | Tag 700\$a:<br>Personal<br>name                            | Tag 752\$a:<br>Personal<br>name                    | Tag 752\$a:<br>Personal<br>name | Tag 752\$x:<br>Personal<br>name |

Example. Henry Esmond : a novel / by Thackeray. Bleak  
House / by Dickens. (from AACR2 2.1G1.)

|                                                                         |
|-------------------------------------------------------------------------|
| Yoshimura, Fuyuhiko<br>吉村 冬彦<br>は<br>Terada, Torahiko<br>寺田 寅彦<br>を見なさい |
|-------------------------------------------------------------------------|

Example of reference card for a pseudonym in NCR 1965 ed.  
Kanji heading written over the Romanized sort key in a double  
 -decker style.

Fig. 1

|           |                                                                 |
|-----------|-----------------------------------------------------------------|
| FO 32     | Satō, Kōiti (1927- )                                            |
| 382       | 小川 正道 (1908 - )                                                 |
|           | 世界の幼児教育 小川正道, 堀沢 一郎, 佐藤浩一著                                      |
|           | 大阪 新興出版社啓林館 昭和 51 (1976)                                        |
|           | 191p 22 cm (幼児教育革新のために各論シリーズ 第1巻<br>幼児教育研究会編)                   |
|           | 付: 参考文献                                                         |
|           | Ogawa, Masamichi I. Hori, Shin'ichirō (1943 - )                 |
| 1062458   | B. Satō, Kōiti (1927 - ) III. Sekai no yōsei kyōiku 1. 幼児<br>教育 |
| FO 32-382 | 2000円                                                           |
| 376. 1    | 国図76-9595                                                       |

Added entry card for third co-author, using NDL printed card.  
 Romanized name at the top does not coincide with the main entry  
 heading in Chinese character in the second line.

Fig. 2

|                             |                     |
|-----------------------------|---------------------|
| 810.1                       | セキネ, マサオ            |
| コ                           | コウズ, ハルシゲ           |
|                             | 高津 春繁               |
|                             | 古代文字の解説 高津春繁 関根正雄著  |
|                             | 東京 岩波書店 昭和39(1964)  |
|                             | v, 302, 17p 地図 19cm |
| 1. セキネ, マサオ 2. コダイ セジノ カイドウ |                     |

Example of an added entry card for second co-author.  
Kana heading at top corresponds to the second author's name  
in Chinese character, underlined in red in the author statement.

Fig. 3

|         |                                                 |
|---------|-------------------------------------------------|
| Y76     | Kukyô de kitaeru siko keihatu                   |
| 2669    | 松本 順 (1917 - )                                  |
|         | 苦境で鍛える自己啓発                                      |
|         | 東京 経林書房 昭和51(1976)                              |
|         | 266p 19cm                                       |
|         | Matumoto, Zyun I. Kukyô de kitaeru siko keihatu |
| 1058947 |                                                 |
|         | 950円                                            |
|         | 76B-2176                                        |

Title added entry card using NDL card.  
Word separation in the romanized title does not affect the filing  
order, because letter-by-letter arrangement is adopted in  
almost all Japanese libraries.

Fig. 4

知られざる地球 Dr. バラードのサイエンス・ドキュメント  
 Robert D. Ballard 著 徳山明〔ほか〕訳  
 東京 福武書店 1984.11  
 381p 29cm  
 Exploring our living plant.の翻訳  
 日本語版監修：市川正巳  
 文献：p364  
 ISBN4-8288-1403-5

84W55501 1. Shirarezaru chikyô a1. Ballard, Robert D.  
 a2. Tokuyama, Akira sl. 地球 ①ME21 ②450

ME21-46

60.07.02

4

070021

40

7000円  
 JP85-25955

Fig. 5



Shirarezaru chikyô

知られざる地球 Dr. バラードのサイエンス・ドキュメント  
Robert D. Ballard 著 徳山明 (ほか) 訳  
東京 福武書店 1984.11  
381p 29cm  
Exploring our living plant. の翻訳  
日本語版監修: 市川正巳  
文献: p364  
ISBN4-8288-1403-5

84W55501 1. Shirarezaru chikyô al. Ballard, Robert D.  
a2. Tokuyama, Akira sl. 地球 ①ME21 ②450

ME21-46 60.07.02 4 070021 102 7000円  
JP85-25955

Fig. 6

シラレザル チキュウ

知られざる地球 Dr. バラードのサイエンス・ドキュメント  
Robert D. Ballard 著 徳山明 (ほか) 訳  
東京 福武書店 1984.11  
381p 29cm  
Exploring our living plant. の翻訳  
日本語版監修: 市川正巳  
文献: p364  
ISBN4-8288-1403-5

84W55501 1. シラレザル チキュウ al. Ballard, Robert D. a2. トクヤマ,  
アキラ sl. 地球 ①ME21 ②450

ME21-46 61.07.15 4 070022 100 7000円  
JP85-25955

Fig. 7

Tokuyama, Akira (徳山明)

知られざる地球 Dr. バラードのサイエンス・ドキュメント

Robert D. Ballard 著 徳山明 (ほか) 訳

東京 福武書店 1984.11

381p 29cm

Exploring our living plant. の翻訳

日本語版監修: 市川正巳

文献: p364

ISBN4-8288-1403-5

84W55501

1. Shirarezaru chikyū a1. Ballard, Robert D.

a2. Tokuyama, Akira s1. 地球 ①ME21 ④450

ME21-46

60.07.02

4

070021

3 4 7

7000円

JP85-25955

### Fig. 8

トクヤマ, アキラ (徳山明)

知られざる地球 Dr. バラードのサイエンス・ドキュメント

Robert D. Ballard 著 徳山明 (ほか) 訳

東京 福武書店 1984.11

381p 29cm

Exploring our living plant. の翻訳

日本語版監修: 市川正巳

文献: p364

ISBN4-8288-1403-5

84W55501

1. シラレザル チキユウ a1. Ballard, Robert D. a2. トクヤマ,

アキラ s1. 地球 ①ME21 ④450

ME21-46

61.07.15

4

070022

2 9 8

7000円

JP85-25955

### Fig. 9

Chikyū (地球)

知られざる地球 Dr. バラードのサイエンス・ドキュメント  
Robert D. Ballard 著 徳山明 (ほか) 訳  
東京 福武書店 1984.11  
381p 29cm  
Exploring our living plant. の翻訳  
日本語版監修: 市川正巳  
文献: p364  
ISBN4-8288-1403-5

84W55501 1. Shirarezaru chikyū al. Ballard, Robert D.  
a2. Tokuyama, Akira sl. 地球 ①ME21 ②450

ME21-46 7000円  
JP85-25955  
60.07.02 4 070021 405

### Fig. 10

チキウ (地球)

知られざる地球 Dr. バラードのサイエンス・ドキュメント  
Robert D. Ballard 著 徳山明 (ほか) 訳  
東京 福武書店 1984.11  
381p 29cm  
Exploring our living plant. の翻訳  
日本語版監修: 市川正巳  
文献: p364  
ISBN4-8288-1403-5

84W55501 1. シラレザル チキウ al. Ballard, Robert D. a2. トクヤマ,  
アキラ sl. 地球 ①ME21 ②450

ME21-46 7000円  
JP85-25955  
61.07.15 4 070022 397

### Fig. 11

## Comparison of UNIMARC and J/MARC tags

An example from AACR 2 2.1G1.

Henry Esmond : a novel / by Thackeray. Bleak House :  
a novel / by Dickens.

UNIMARC tags for above example

Tag 200\$a = Henry Esmond = Title proper 1 = TP1

Tag 200\$e = a novel = Related title 1 = RT1

Tag 200\$f = by Thackeray = Statement of responsibility 1 = SR1

Tag 200\$c = Bleak House = Title proper 2 = TP2

Tag 200\$e = a novel = Related title 2 = RT2

Tag 200\$f = by Dickens = Statement of responsibility 2 = RS2

J/MARC tags for above example

TP1 = Tag 251\$A == Tag 551\$A = kana AC1 = Access point of TP1  
== Tag 551\$X = romanized AC1

RT1 = Tag 251\$B (no access point)

SR1 = Tag 251\$F == Tag 751\$A = kana AC1 = Access point of SR1  
== Tag 751\$X = romanized AC1

TP2 = Tag 252\$A == Tag 552\$A = kana AC2 = Access point of TP2  
== Tag 552\$X = romanized AC2

RT2 = Tag 252\$B (no access point)

SR2 = Tag 252\$F == Tag 752\$A = kana AC2 = Access point of SR2  
== Tag 752\$X = romanized AC2

Fig. 12



1986

PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS: PROBLEMS AND SOLUTIONS.  
Organized by the Section on Library Services  
to Multicultural Populations and the Section  
on Information Technology.

Tokyo, August 21-22

71-MCULTP-10-E  
INF

An Universal Coding System

for Multi-Lingual Environment

by

S.S. TSENG, C.C. HSIEH;  
Jack K.T. HUANG, C.T. CHANG  
and C.C. YANG

TAIPEI, CHINA

KEYWORD(S):

For internal use only:

To be brought to meeting no:

Thursday 21 August '86

Remarks:

to be presented orally ☒

for distribution only ☐

194

096

多言語を処理するソフトウェア環境における3バイト、2バイト、1バイトのグラフィック文字のコーディング・システムについて述べる。

このシステムの構造は1980年に開発されたCCCII(Chinese Character Code for Information Interchange)をベースとしており、最初の版は1983年に完成した。CCCIIの第3版は、ISO-646やISO-2202に基づいたすべての3バイト、2バイト、1バイトのコーディング・システムにも充分対応することができる。文字セットは台湾、日本、韓国、中国のそれぞれの文字を包含している。第3版には標準的な42,000の漢字があり、さらに11,000以上の変化した形の漢字が含まれており、JIS C6226、GB 2312も入っている。これらの文字セット間での変換機構がこのシステムで提供される。

CCCIIの3バイト方式は3、2、1各バイトの文字セット間での相互対応に都合の良い構造となっている。CCDB(Chinese Character Data Base)と呼ぶ文字の属性を結合するデータベースがあり、相互に属性をアクセスできる。たとえば中国の漢字から、CCDBの索引機能によって対応する日本語の漢字を呼び出せるのである。

CCCIIはISO-646およびISO-2202に基づく3つの7ビットバイトより成り、文字のための領域としては94×93×94すなわち合計821,748の文字とデータ処理のための制御文字を収容できる。先頭(第1)バイト(B1)はセクション中のポジション(位置)、第2バイト(B2)は1つのプレーン(Plane-面)の中のセクション、第3バイト(B3)はプレーンを表わしている。CCDBは漢字のためのデータベースであり、必要なソフトウェア群に関連した属性のデータベースである。このようにCCCIIとCCDBは、7万以上の漢字に最適であるだけでなく、多くの中国語や準中国語ともいべき日本語、韓国語等のいわゆる東洋言語に最適である。

国際間の情報処理システムにとって多言語コーディング・システムが必要なことはいうまでもない。CCCIIとCCDBはこのような仕事に秀れた環境を提供する。CCDBは次の3つの主要機能をもっている。(1)漢字の構成、読みあるいは発音記号、漢字のドット・マトリックス構成、他のコーディング・システムのコードなどを提供する。(2)漢字のフォントと読み方の間の変換テーブルを制御し、検索する機構を提供する。(3)複数バイトの種々のコーディング・システム間の変換機能を提供する。例えば3バイトのCCCIIから2バイトのJISC-6226へ、あるいはその逆など。

ユニバーサルなコーディング・システムをめざしてCCCIIの構成に基づいたCCCII機構が設計された。CCCII機構システムは次の3つの要素で構成されている。(1)核(Kernel): CCDBは実はこの核なのである。適切なルーチンの呼び出しによって、CCDBの情報がアクセスされる。(2)変換ユニット: これはCORE(CODE RECEIVER—コード入力)とCOTRA(CODE TRANSMITTER—コード変換出力)からなり、CCDBとデータ通信装置との間のインターフェイスの役割を果たす。(3)コントローラ(制御部): コントローラはCCCII機構の中心となるコントロール・プログラムである。CCCII機構は端末やPC(パーソナルコンピュータ)やホストコンピュータに実装されることになるのであろうし、種々のアプリケーションのいろいろな部分で、それぞれの目的に応じた機能を提供することになる。したがってコントローラはアプリケーションに依存するものである。CCCII機構は単なる机上の伝説ではなく、すでに製品化の準備が進められており、台湾では数社が開発中なので、今年末には市場でお目にかかれるものと期待されている。

CCCIIにはデータ処理のため最も大きなコーディン容量(主要文字領域)がある。CCCIIの3バイト構造は、漢字の正規の形、簡略形、変化形の関係を維持し、日本語の仮名や韓国語のハングルをはじめほとんどすべてのアルファベティックな文字もそのユニークなコーディング・システムの下で処理することができる。多言語を処理する場合でも複数の文字セットを使用する必要がなく、エスケープ・シーケンスも不要であり、単一のコーディング・システムで済むのである。

1986年末あるいは1987年初めには7万字を越える文字数を取容するCCCIIの第4版を出す予定である。国際情報交換だけでなく、多言語ソフトウェアの開発にも有効と考える。



## AN UNIVERSAL CODING SYSTEM FOR MULTI-LINGUAL ENVIRONMENT

S.S. TSENG, C.C. Hsieh,  
JACK K.T. HUANG, C.T. CHANG  
and C.C. YANG

CHINESE CHARACTER ANALYSIS GROUP  
COUNCIL FOR CULTURAL PLANNING  
AND DEVELOPMENT  
TAIPEI, REPUBLIC OF CHINA

### Abstract

In this paper. An interchange coding system of 3-byte/ 2-byte/ 1-byte graphic symbols for multi-lingual software environment will be presented. The structure of this system is based on CCCII (Chinese Character Code for Information Interchange) developed in 1980. The first version of this system was completed in 1983.

The version 3 of CCCII is flexible enough to adopt any existing 3-byte/ 2-byte/ 1-byte coding systems that are based on ISO-646 and -2022. At this moment, it has space to accomodate more than 90 7-bit bytes standards. Therefore, most of the published 7-bit byte standards may be integrated into this system without system modification. For Oriental languages, the character sets collected include those used in R.O.C., Japan, Korea and mainland China. There are 42,000 regular Chinese characters in this version 3 character set. Besides, the variations for the regular characters collected are beyond 11 thousands. JIS C 6226 and GB 2312 character set are also located in the version 3 of CCCII. A mechanism of interchange among these character sets is provided within the system.

The 3-byte structure of CCCII provides a convenient mechanism for cross mappings among different 3-byte/ 2-byte/ 1-byte character sets. Used in conjunction with the attribute data base of characters, the CCDB (Chinese Character Data Base), this system provides attribute cross-reference access features among different 3-byte/ 2-byte/ 1-byte standards. For example, an user can have access to Japanese Kanji through the index mechanism of CCDB of corresponding Chinese character.

At present, the CCCII system structure is fixed but its symbol set is still open. We expect that in the later part of 1986 or in early 1987 the characters collected in the version 4 of CCCII will exceed 70 thousands. This CCCII coding system is not only beneficial for international information interchange, but also valuable for developing multi-lingual software in tomorrow's computer system.

## Introduction to Multi-lingual Environment

The multi-lingual environment is based on CCCII coding system [1] and CCDB [2] which issued in 1980 and 1982 respectively. CCCII is a three 7-bit-byte code based on ISO-646 and -2022, the graphic area of this system has  $94 \times 93 \times 94$ , totally 821,748 coding positions and 94 control codes for data processing. The first byte, B1, denotes the positions within a section, the second byte, B2, denotes the sections of a plane, and the third byte, B3, denotes the planes, as shown in figure 1. CCDB is a data base for Chinese characters and their associated attributes with a set of necessary software tools. The software structure and index mechanism of CCDB are shown in figure 2-3 respectively. Under those structures of CCCII and CCDB, it is not only adequately used for more than 70 thousands Chinese characters, but also adequately used for the Oriental languages which use a lot of Chinese characters or Chinese-likely characters, such as in Japan and Korea [3].

In fact, we need a multi-lingual coding system for international information processing systems [4], here with the CCCII and CCDB provide an excellent environment for the work. The advantages are explained as follows:

- (1) The implementation of CCCII is shown in fig 4, 94 planes divide into 16 layers, each layer has 6 planes except layer 16 has only 4 planes. The first section of every plane is reserved for CCCII control code which will explain later, section 2 to 15 of the first plane of each layer are reserved for user area except layer 13, 14, 15, and 16. Please see Fig. 5 and 6.
- (2) The first layer allocates the regular form of Chinese characters. The second layer through layer 12 allocate the variant forms of Chinese characters and the simplified forms are treated as variant forms which are allocated in layer 2 only.
- (3) We treat the 6,798 Kanji of JIS C 6226 as variant forms of chinese characters and allocate them in layer 13. The other symbols of JIS C 6226 then allocated in the section 2 to 15 of layer 13. By the mechanism of CCDB, the Japanese Characters can be coded in CCCII structure in such a way, and we had done it already.
- (4) Layer 14 is reserved for Korean characters in the same way as allocate the Japanese characters in layer 13.
- (5) Plane 85, the first plane of layer 15 is reserved for characters of minority Chinese languages such as Mongolian, Tibetan, Manchurian, Moslem, etc.
- (6) Plane 86 through plane 94 are reserved for the other languages

of the world. There are adequately enough coding spaces and convenience mechanisms for an universal language coding system.

- (7) The control codes of CCCII include two parts, the first part of them are used for the extension techniques by means of escape sequence[5]. The second part of them are used for edition or other applications of text files. These codes keep B3 = blank, B2 = 33 as control code indicator, practically, only B1 is used. We have assigned 27 control codes as shown in fig. 7, it can be increased by practice requirement. Those codes are grouped as follows:

(a) Text Separators:

DPA ( Delimiter between Parts )  
DCH ( Delimiter between Chapter )  
DSE ( Delimiter between Section )  
DBL ( Delimiter between Blocks / Subsections )  
DPR ( Delimiter between Paragraphs )  
DST ( Delimiter between Sentences )  
DWO ( Delimiter between Words )  
DTI ( Delimiter between Titles )  
DCU ( Delimiter between Chuns (rolls) )  
DVO ( Delimiter between Volumns )  
DPG ( Delimiter between Pages )  
DLN ( Delimiter between Lines )

(b) Typesetting Effectors:

SCH ( SKip one Character )  
SLN ( SKip one Line )  
CHS ( Character Scaling )  
CSR ( Character Scale Recovery )  
CFR ( Character-Font Recovery )  
CDR ( Character-Display Reverse )  
CHL ( Character High-light )  
HLR ( High-light Recovery )

(c) Code-Format Switchers:

PSL ( Plane-Shift, Locked )  
PSR ( Plane-Shift, Recovery )  
PSU ( Plane-Shift, Unlocked )  
SSL ( Section/plane-shift, Locked )  
SSU ( Section/plane-shift, Unlocked )  
HOM ( Home )

Now, we will introduce a mechanism of universal coding system as below.

## The mechanism for universal coding system

### 1. Introduction to the CCDB

Chinese Character Data Base (CCDB), shown as fig. 2, is a data base together with necessary software that was developed based on the structure of CCCII in order to make the application of CCCII easy and efficient. CCDB has three main functions:

- (1) To provide Chinese character constructions, phonetic transcription, dot matrix configurations of Chinese characters, various correlated input methods and various existed corresponding Chinese character coding systems [5], please see figure 8.
- (2) To provide controlling and searching mechanism of transcription table between corresponding lexicographic character fonts and phonetic symbols.
- (3) To provide the interchanging function among various existed corresponding multi-bytes coding systems, for example, from 3-byte CCCII to 2-byte JIS C 6226 or from 2-byte JIS C 6226 to 3-byte CCCII, etc.

The CCDB file structure is mainly organized by two portions: data files and indexing files. The data files subdivided into 16 layers. All files are grouped into seven file-groups as described below.

- (1) G1 (files of layer 1): designates the data files of regular form of Chinese characters.
- (2) G2 (files of layer 2): designates the data and searching files of simplified forms of variant forms of Chinese characters.
- (3) G3 (files of layer 3-12): designates the data and searching files of other variant forms of Chinese characters appearing in layer 1. Currently, only five selected variant forms are vertically allocated from layer 3 to 7.
- (4) G4 (files of layer 13): the data and searching files of Japanese Kanas and 6,798 Kanji of JIS C 6226 are allocated in layer 13.
- (5) G5: Inverted lists of Chinese lexicographic fonts, phonetic symbols, and the other attributes of Chinese Characters.
- (6) G6: Indexed tables of various Chinese character input methods, e.g., 3-corn code, Dragon input method, etc.
- (7) G7: Cross-reference tables of various existed relevant coding



systems.

The Chinese Characters stored in CCDB are hired R94, a 16-bit condensed code of CCCII. The file groups separated by a layer pointer LP. To calculate R94 and LP, CCDB provide formulas as

$$R94 = [(B3-33) \bmod 6] * 94 * 94 + (B2-34) * 94 + (B1-33), \text{ and} \\ LP = (B3-33) / 6.$$

The searching mechanism of CCDB uses three different addressing modes: (see Figure 3)

- (1) Direct addressing: If  $LP=0$ , the R94 code is the storing address of the regular form of Chinese Characters in layer 1 in the environment of CCCII.
- (2) Indirect addressing: If  $LP=1$ , the R94 code is the address of the storing address of the simplified form of Chinese characters in layer 2. If  $LP=12$ , R94 code is the indirect address of Japanese Kanas and Kanji of JIS C 6226 allocated in layer 13 in the environment of CCCII. If  $LP=13$ , the R94 code is the indirect address of KIPS or KSC-5619 Korean characters which to be allocated in the environment of CCCII.
- (3) Indirect-displacement addressing: If  $LP=2$  to 11, the R94 code is the address of a Base, denoted by B, and the storing address of the variant forms of Chinese characters in layer 3 through layer 12 in the environment of CCCII are calculate by the formula  $m = B + LP - 2$ , where LP must less than or equal to the upper boundary of LP (L). The upper boundary of LP and B are different from character to character, so it should be pre-determined when the character was stored.

The above mentioned three addressing modes are also applicable to currently reserved 85-94 planes of CCCII. The CCDB provides this mechanism for multi-lingual data procesing as an universal coding system as long as those languages were coded within the scope of CCCII. One important and common feature of CCCII and CCDB we would like to point out that the both mentioned systems are open systems in nature, subject to increment of appending of relevant searching and data files whithout further modification of the systems.

## 2. Code-format Switchers:

Since CCDB is based on 3-byte CCCII. For up-grade the efficiency of data communication and storage, the 3-byte code can be transfer to 2-byte or 1-byte code in accordance with the usage of control codes defined by CCCII[6]. Those set of control codes are called code-format switchers. CCDB provides the control processing algorithm of code-format switchers as shown in Appendix A.

## 3. Introduction lto the CCCII-machine:

For the goal of implementation of universal coding system, the CCCII-machine was designed that based on the structure of CCCII. The CCCII-machine system constructed by three main components, namely; (1) Kernel, (2) Transmission Unit and (3) Controller. Please see figure 9.

- (1) Kernel: CCDB is the kernel in fact. By means of the proper routine calls, the information of CCDB will be accessed. Some flags tell the states of activities.
- (2) Transmission unit: The transmission unit consists of the CORE (Code REceiver) and COTRA (Code TRAsmitter). It serves as an interface between CCDB and data communication devices. The CORE and COTRA are constructed by specific data structure and routines. Under the proper routine calls, the CORE and COTRA will be activated. Please refer to figure 10 and 13.
- (3) Controller: Controller is the principal control program of CCCII-Machine. Because CCCII-Machine may implement on terminal, PC, or host computer, the controller then provide different functions in the different places of different applications. So the controller is application dependence, we will not go far about it then.

#### 4. CORE (Code REceiver):

Refer to figure 10, the CORE recieves a character stream sending from the data transmission device. The character stream usually is an mixture including the control codes of ISO-646, the control codes of CCCII, the 3-byte and/or 2-byte and/or single byte codes of CCCII. The responsibilities of CORE are to pick up those who and give them the proper treatment and process, especially, the single byte or 2-byte codes should be transfer back to 3-byte representations. The data structure of CORE is explained as follows:

- RI: 7-bit input register.
- RB3: 7-bit register stores B3 of CCCII.
- RB2: 7-bit register stores B2 of CCCII.
- RB1: 7-bit register stores B1 of CCCII.
- SD3: A stack to keep the B3 default values for PSL/PSU/PSR control codes under 2-byte condition.
- SP: Stack point for SD3.
- RD3 and RD2: 7-bit registers store B3 and B2 default values for SSU code under 1-byte condition.
- S: State register, value 0 indicates the 3-byte condition, value 1 indicates the 2-byte condition and value 2 indicates the 1-byte condition.
- C: Modula 3 register for byte position control, if 0 indicates next byte into RI is B3.



- 1 indicates next byte into RI is B2.
- 2 indicates next byte into RI is B1.
- F: Register for error status.

The routine componets of CORE composed of a main control routine and a group of correlated subroutines. The flowchart of main control routine is shown in figure 12. Main control routine is an infinite-loop by nature. Once controller starts, the main control routine of CORE will ceaselessly operate.

The main control routine is also served as an interrupt-driven routine. It is liable to idle after accepted a new byte and processed. It will wake-up until next byte enters the RI. To stop the main control routine of CORE must be operated by means of system break. The details of design of idle, wake-up and break subject to implement-dependent manufacturers' and users' wish.

Error-handling functions are included in each subroutine. In case of error occurs, the status of CORE will be handled by F register of CORE, the the controller will take care after. A set of routines which wrote in pascal is shown in Appendix B.

#### 5. COTRA (CODE TRANsmitter):

Same as CORE, the COTRA is constructed by two components; the data structure and the routines as shown in figure 13. In the data structure, except the S is defined slight differently from CORE, the definitions of RB3, RB2 and RB1 are the same as in CORE.

S is designated to control the number of bytes of CCCII to transmit out. When Controller calls COTRA to be requested to transmit CCCII. Functions of S are as follows:

- (1) If S=0, then transmit all the RB3, RB2 and RB1.
- (2) If S=1, then transmit the RB2 and RB1.
- (3) If S=2, then transmit the RB1.

It is necessary to utilize routine calls to modify the contents of S. If the content of S is changed, COTRA will transmit corresponding control codes and parameter (if needed), as shown in figure 14.

There is no main control routine for COTRA routines, because COTRA is totally passive and completely controlled by Controller. A set of programs for COTRA routines which wrote in PASCAL is shown in Appendix C.

## Conclusion

The CCCII-machine is not merely a hypothetic system on a drawing board, with its complete structure design and necessary detailed data it is ready for manufacturing. The CCCII-machine can be implemented on terminals, PC, or host computers. There are several manufacturers in Taiwan already taking place in developing the CCCII-machine. It is expected to see them on the market at end of this year, if every thing goes well.

Since CCCII provides the largest coding spaces (821,748 characters) available for data processing purpose. The 3-byte structure of CCCII allows for the multi-lingual data processing. The internal logic of 3-byte CCCII makes possible the linkage of regular, simplified, variant forms of Chinese character, and others such as Japanese Kana and Kanji, Korean Hanguel, and almost all alphabetic characters to be processed all under one unique coding system. To avoid of multiple character sets in processed all under one unique coding system. To avoid of multiple character sets in processing of mulit-lingual environment, invoke escape sequences between them, and thus arbitrarily partitioning a unitary coding system.

The mechanism of CCCII-machine offers a single, coherent coding system and extends the applications limit for multilingual data processing in all fields, naturally, the application of library automation is included.

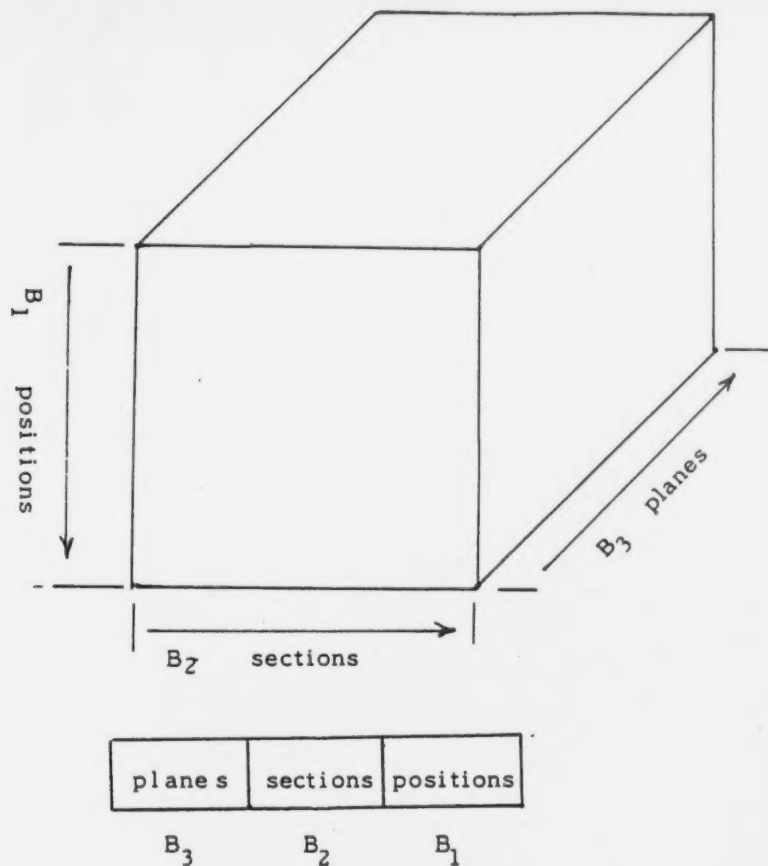


Figure 1. Coding structure of CCCII

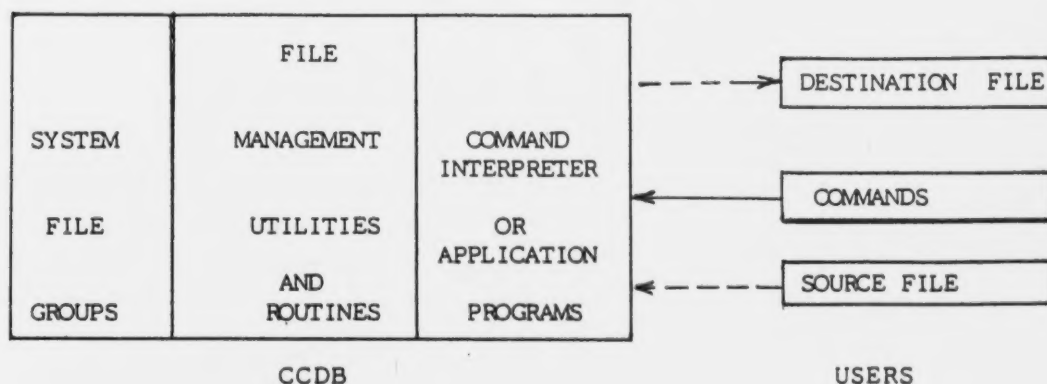


Figure 2. Software Structure of CCDB

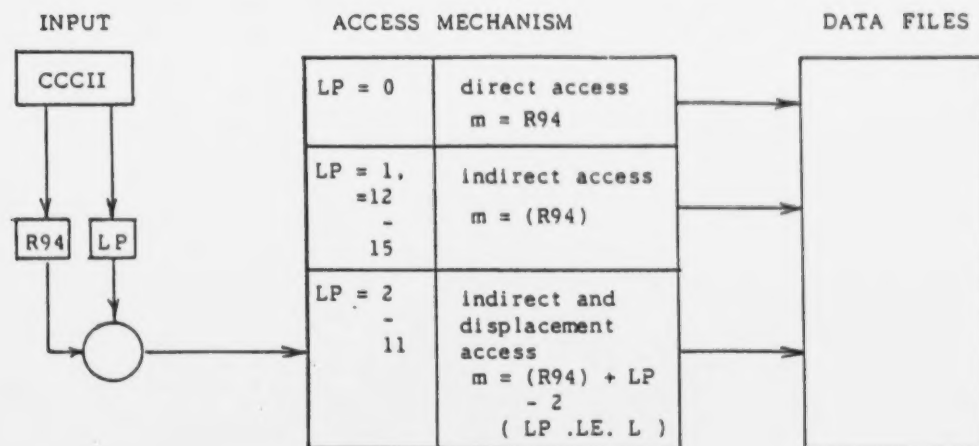


Figure 3. THE ACCESS MECHANISM OF CCDB

LP: Layer indicator, from 0 to 15

L: The upper boundary of LP

m: The effective address of the character to be access

R94: Condensed code of CCCII

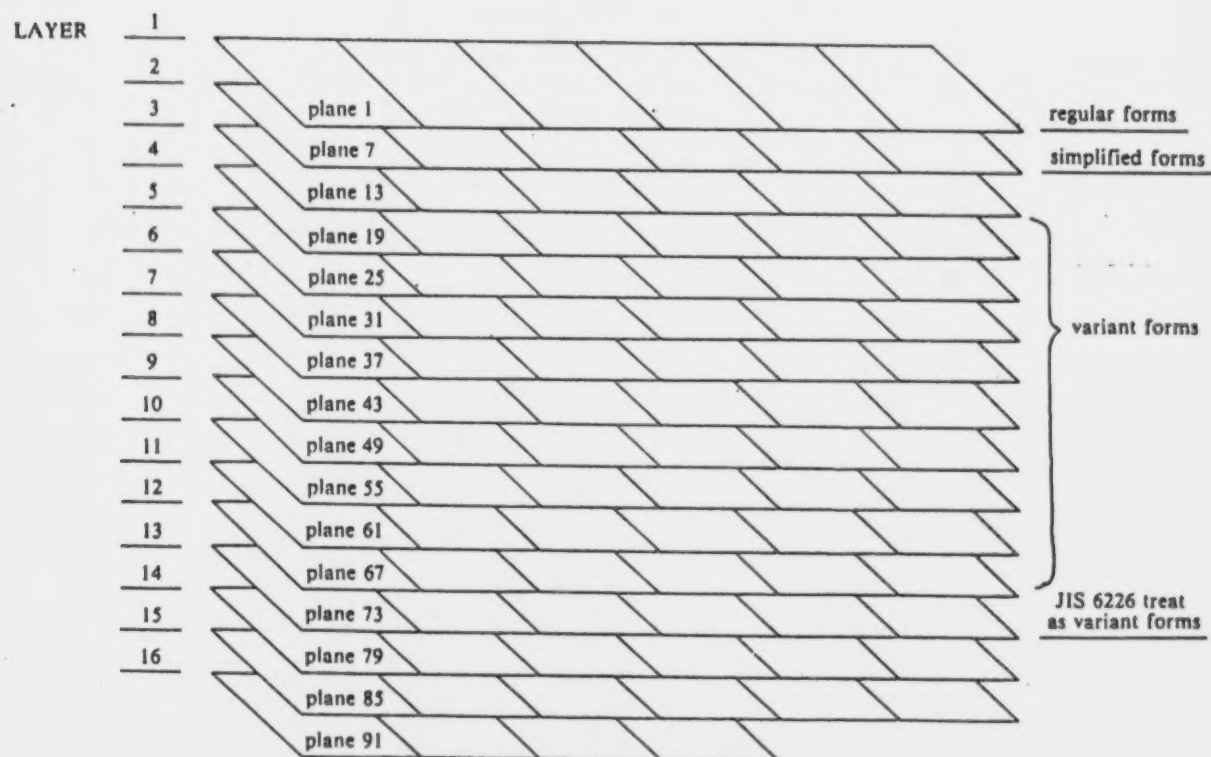


Fig. 4. Structure and Allocation of CCCII

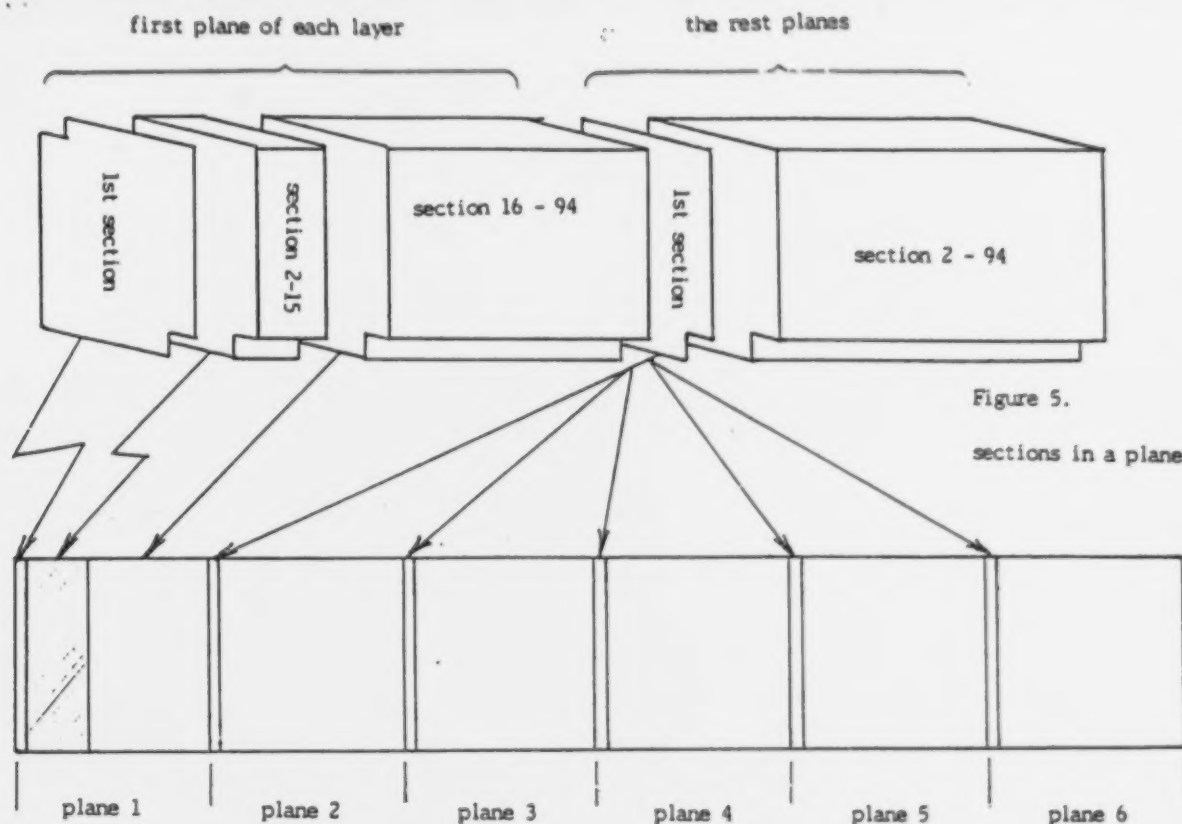


Figure 6. a layer is composed of 6 planes  
 the 1st section of each plane is reserved for the control codes of CCCII  
 section 2 - 15 of the 1st plane of layer 1 - 12 are reserved for user area

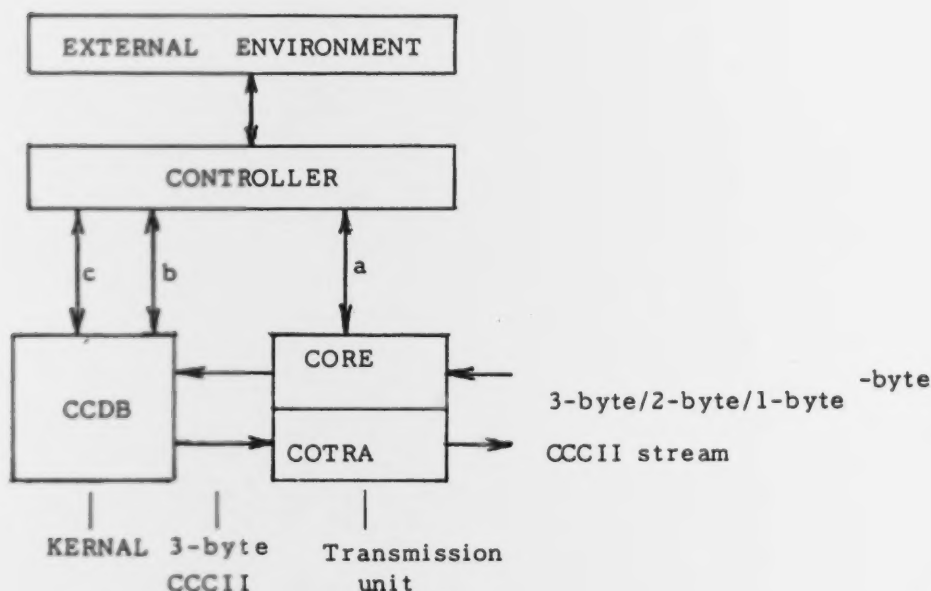
| R/C | 2    | 3 | 4   | 5   | 6 | 7    |
|-----|------|---|-----|-----|---|------|
| 0   | //// |   | DWO | SCH |   | PSL  |
| 1   |      |   | DST | SLN |   | PSR  |
| 2   |      |   | DPR | CHS |   | PSU  |
| 3   |      |   | DBL | CSR |   | SSL  |
| 4   |      |   | DSE | CFS |   | SSU  |
| 5   |      |   | DCH | CFR |   |      |
| 6   |      |   | DPA | CDR |   |      |
| 7   |      |   |     | CHL |   | HOM  |
| 8   |      |   | DLN | HLR |   |      |
| 9   |      |   | DPG |     |   |      |
| 10  |      |   | DVO |     |   |      |
| 11  |      |   | DCU |     |   |      |
| 12  |      |   | DTI |     |   |      |
| 13  |      |   |     |     |   |      |
| 14  |      |   |     |     |   |      |
| 15  |      |   |     |     |   | //// |

Figure 7. CONTROL CODES OF CCCII

|       |     |     |     |     |     |     |     |     |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|
| CCCII | RAD | STK | EXP | VFM | PHC | FNT | ETL | OTC |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|

CCCII: Chinese Character Code for Information Interchange  
 RAD: Radical which a character belongs to  
 STK: Total stroke count  
 EXP: Component expression of a character  
 VFM: Variant forms of a character  
 PHC: Phonetic symbols of a character's pronunciation  
 FNT: Dot matrix representation of the FONT of a character  
 ETL: External or Indexing code of a character  
 OTC: Codes of different coding system

Figure 8. The Attributes Provide by CCDB



- The routine calls from contrdler to CORE/COTRA and STATUS from CORE/COTRA back to controller
- The routine calls from controller to CCDB and the STATUS from CCDB back to controller
- Data channels

Figure 9. The organization of CCCII - MACHIME



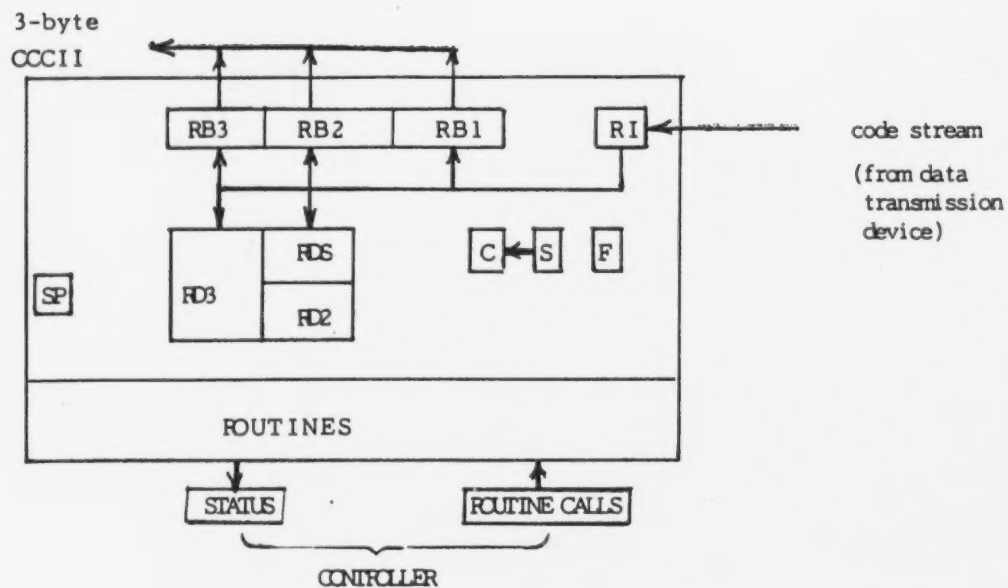


Figure 10. The Structure of CORE

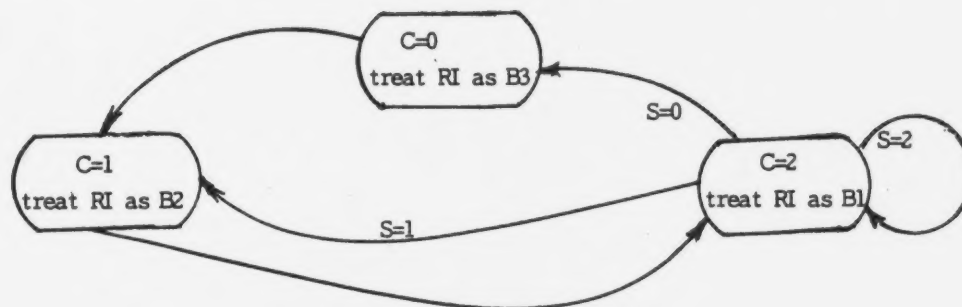


Figure 11. The functions of C and S

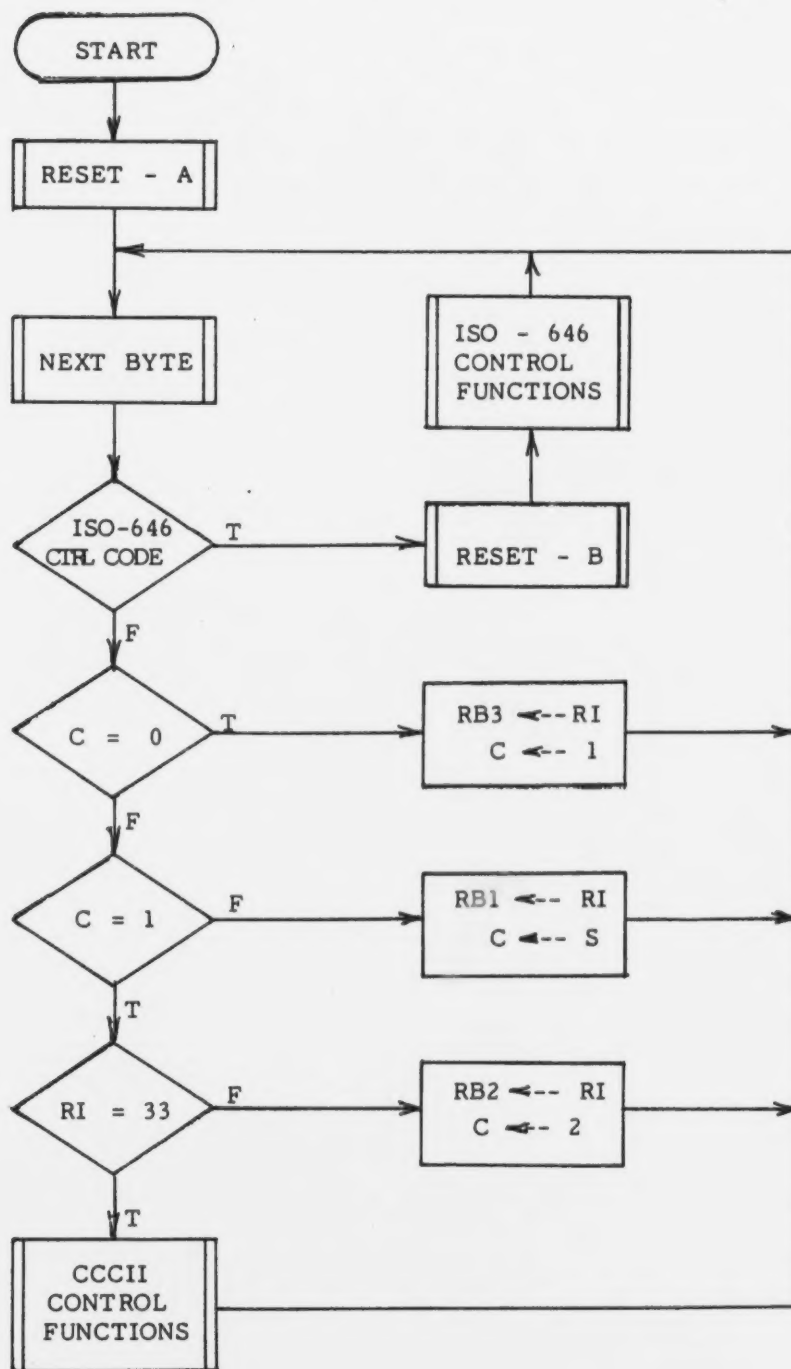


Figure 12. The main control routine of CORE

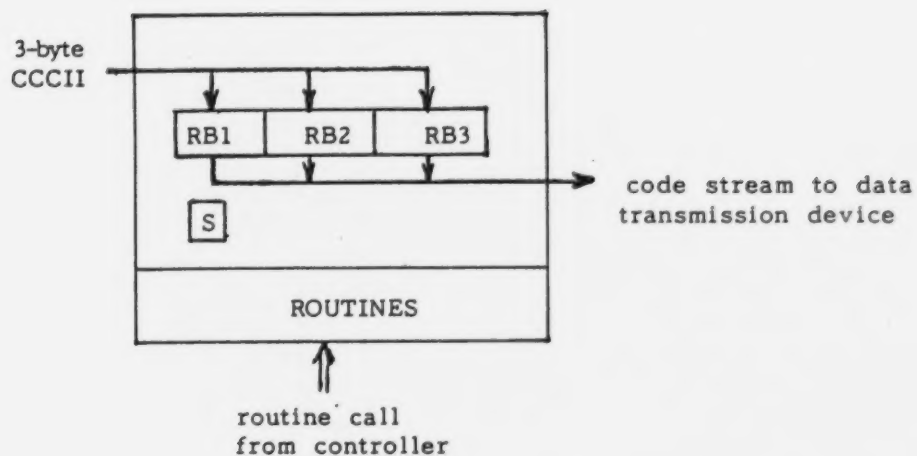


Figure 13. The structure of COTRA

| S       | OUTPUT                   | DESCRIPTION       |
|---------|--------------------------|-------------------|
| 0 --> 1 | PSL / PSU with parameter | 3-byte --> 2-byte |
| 0 --> 2 | SSL with parameter       | 3-byte --> 1-byte |
| 1 --> 0 | HOM                      | 2-byte --> 3-byte |
| 1 --> 2 | SSL with parameter       | 2-byte --> 1-byte |
| 2 --> 0 | ESC 2/4 4/1              | 1-byte --> 3-byte |

Figure 14. The relation between control code and S

## Reference

- [1] The Chinese Character Analysis Group, "Symbol and Character Tables of Chinese Character Code for Information Interchange", May, 1983.
- [2] The Chinese Character Analysis Group, "The design of a cross-reference database for Chinese Character Indexing", Apr., 1980.  
"Chinese Character Data Base, its Design, Implementation, and Application", at ASIS-82, October 19, 1982.
- [3] The Chinese Character Analysis Group, "A Multi-lingual Coding System based on CCCII", 2nd Asian-Pacific Conference on Library Science, May 20, 1985.
- [4] Karen Smith-Yoshimura & Alan Tucker, "RLIN East-Asian Character Code and RLIN CJK Thesaurus", 2nd Asian-Pacific Conference on Library Science, Seoul, Korea, May 20, 1985.
- [5] The Chinese Character Analysis Group, 中國文字資料庫, 2nd edit., May, 1985, Taipei.
- [6] C. C. Hsieh, et. al., "對 CCCII 控制碼的說明", CCCII 維護技術資料, 文建會, March 13, 1986, Taipei.

## Appendix A

\* PSL ( Plane-shift, locked ) : B1=112

To set the B3 ( plane ) as default value, all the following CCCII's will have the same B3 and thus omitted. This will produce 2-byte ( B2 and B1 ) CCCII's. PSL must be followed by 2-byte parameter, in which the first byte is ignored and the second byte is used as the plane indicator. The control function of PSL is explained as follow :

|                                                                                         |              |           |                                                      |
|-----------------------------------------------------------------------------------------|--------------|-----------|------------------------------------------------------|
| {[B3],B2,B1;}                                                                           | [B3],33,112; | X,B3';    | {B2,B1}                                              |
| 3-byte CCCII<br>string or 2-<br>byte CCCII<br>string with<br>old default<br>value of B3 | PSL          | parameter | 2-byte CCCII<br>string with<br>KB3' default<br>value |

\* PSR (Plane-shift, Locked): B1=113

Pop the last old default value of B3 from stack, if stack is bank, the PSR will be not reacted. The PSR has no parameter follows after, and it is accepted only under the case of 2-byte condition, if under the case of 3-byte condition, the PSR will be rejected.

\* PSU (Plane-shipt, unlocked): B1=114

Similar to PSL. The difference is that PSL will change the B3 permanently (or until another PSL with different plane value). But the PSU will only change the following character. The PSU is accepted only under the case of 2-byte condition, if under the case of 3-byte condition, the PSU will be rejected. The function of PSU is explained as follows:

|                                                    |         |           |                                 |                                                 |
|----------------------------------------------------|---------|-----------|---------------------------------|-------------------------------------------------|
| {B2,B1; }+                                         | 33,114; | X,B3';    | B2,B1;                          | {B2,B1}+                                        |
| 2-byte<br>CCCI<br>stream<br>with<br>default<br>B3+ | PSU     | parameter | 2-byte<br>CCCI<br>(with<br>B3') | 2-byte<br>CCCI<br>stream<br>with<br>default B3+ |

\* SSL (Section/Plane-shift, Locked): B1=115

Use to set the default values of B3 and B2. All the following CCCII's. SSL is applicable to 3-byte and 2-byte condition, the SSL

must be followed by a 2-byte parameter, the first byte indicates the default value of B2, and the second byte indicates the default value of B1. The functions of SSL is explained as follows:

|                                                           |              |           |                                                                 |
|-----------------------------------------------------------|--------------|-----------|-----------------------------------------------------------------|
| {[B2],B2,B1}                                              | {B3},33,115; | B2',B3':  | {B1}                                                            |
| 3-byte CCCII<br>stream with<br>old default<br>value of B3 | SSL          | parameter | 1-byte CCCII<br>stream with<br>default B3'<br>and default<br>B2 |

\* SSU (Section/Plane-shift, Unlocked): B1=116

Similar to SSL, but only react on the character that followed after. SSU only applies under 2-byte condition, SSU will be rejected if under 3-byte condition. The functions of SSU is explained as follows:

|                                                    |        |           |                                           |                                                    |
|----------------------------------------------------|--------|-----------|-------------------------------------------|----------------------------------------------------|
| {B2,B1;}+                                          | 33,116 | B2',B3'   | B1;                                       | {B2,B1;}+                                          |
| 3-byte<br>CCCI<br>stream<br>with<br>default<br>B3+ | SSU    | parameter | 1-byte<br>CCCI<br>(with<br>B3'<br>and B2' | 2-byte<br>CCCI<br>stream<br>with<br>default<br>B3+ |

\* HOM (Home): B1=119

HOM is for deleting the defined B3 default value, to recover back to 3-byte or 2-byte, once the CCCII 3 bytes is reduced to single byte, it will need the ESC 2/4 4/1 to go back to 3 bytes codes.

Notes:

1. For all CCCII control codes, B3 is not active and B2=33.
2. All the symbol used above, their definitions are as follows:
  - , Used to delimit the byte, not appear in the code stream.
  - ; Used to delimit the CCCII code, not appear in the code stream.
  - [ ] Used to express B3 is not exist under 2-byte condition.
  - { } Used to express CCCII stream, as the number of CCCII codes = 0, 1, 2, .....;



X Used to express the byte value is 34-126.

B3' and B2' Used each expresses B3 and B2 default value respectively.

{ }+ Used to express 2-byte CCCII stream with specified default B3+.

## Appendix B

```
procedure F_RESET;  
begin  
  F:=NO_ERROR  
end;
```

```
procedure RESET_A;  
begin  
  S:=0;  
  C:=0;  
  SP:=BOTTOM;  
  F_RESET  
end;
```

```
procedure RESET_B;  
begin  
  if C<>S then F:=IMCOMPLETE_CCCII;  
  C:=S  
end;
```

```
procedure RESET_C;  
begin  
  C:=S;  
  F:=PARAMETER_ERROR  
end;
```

```
procedure NEXT_BYTE;  
begin  
  repeat WAIT_LOOP { CORE idle }  
  until EVENT(RI) { next byte into RI }  
end;
```

```
procedure ACCEPT_BYTE;  
begin  
  NEXT_BYTE;  
  if RI<33 or RI>126  
  then F:=BYTE_ERROR  
  else F:=NO_ERROR  
end;
```

```

procedure PSL;
begin
  ACCEPT_BYTE;
  ACCEPT_BYTE;
  if F=NO_ERROR
    then
      begin
        RB3:=RI;
        SD3[SP]:=RI;
        SP:=SP+1;
        S:=1;
        C:=1
      end
    else RESET_C;
end;

```

```

procedure PSR;
begin
  if SP>BOTTOM
    then
      begin
        case S of
          0 : F:=PSR_REJECT;
          1 : begin
              SP:=SP-1;
              RB3:=SD3[SP];
              C:=S;
              F_RESET
            end
        end
      end
    else F:=STACK_EMPTY
end;

```

```

procedure PSU;
begin
  ACCEPT_BYTE;
  ACCEPT_BYTE;
  if F=NO_ERROR
  then
    begin
      RB3:=RI;
      ACCEPT_BYTE;
      if F=NO_ERROR
      then
        begin
          RB2:=RI;
          ACCEPT_BYTE;
          if F=NO_ERROR
          then
            begin
              RB1:=RI;
              write(RB3,RB2,RB1);
              RB3:=SD3[SP-1]
            end
          else F:=PSU_FAIL
          end
        end
      else F:=PSU_FAIL
      end
    end
  else F:=PSU_FAIL;
  C:=S
end;

```

```

procedure SSL;
begin
  ACCEPT_BYTE;
  if F=NO_ERROR
  then
    begin
      RB2:=RI;
      RD2:=RI;
      ACCEPT_BYTE;
      if F=NO_ERROR
      then
        begin
          RB3:=RI;
          RD3:=RI;
          S:=2;
          C:=2
        end
      else RESET_C
      end
    end
  else RESET_C
end;

```

```

procedure SSU;
begin
  ACCEPT_BYTE;
  if F=NO_ERROR
  then
    begin
      RB2:=RI;
      ACCEPT_BYTE;
      if F=NO_ERROR
      then
        begin
          RB3:=RI;
          ACCEPT_BYTE;
          if F=NO_ERROR
          then
            begin
              RB1:=RI;
              write(RB3,RB2,RB1);
              RB3:=RD3;
              RD2:=RD2
            end
          else F:=SSU_FAIL
          end
        end
      else F:=SSU_FAIL
      end
    end
  else F:=SSU_FAIL
  C:=S
end;

```

```

procedure HOM;
begin
  RESET_A
end;

```

```

begin
  ACCEPT_BYTE;
  if F=NO_ERROR
    then
      begin
        case RI of
          33 : ..... ;
            :
            :
          112 : PSL;
          113 : PSR;
          114 : PSU;
          115 : SSL;
          116 : SSU;
            :
          119 : HOM;
            :
            :
          126 : .....
        end
      end;

procedure ESC_SEQ;
begin
  NEXT_BYTE;
  if RI=36
    then
      begin
        NEXT_BYTE;
        if RI=65 then RESET_A
          else .....
        end
      else .....
    end;

procedure ISO646_CONTROL;
begin
  if RI=127
    then DELETE
    else
      begin
        case RI of
          0 : NULL;
            :
            :
          27 : ESC_SEQ;
            :
            :
          32 : SPACE
        end
      end;
end;

```



```

procedure CORE;
begin
  RESET_A;
  repeat
    NEXT_BYTE;
    if RI<=32 or RI=127
    then
      begin
        RESET_B;
        ISO646_CONTROL
      end
    else
      begin
        case C of
          0 : begin
              RB3:=RI; C:=1
            end;
          1 : if RI=33
              then CCCII_CONTROL
              else
                begin
                  RB2:=RI; C:=2
                end;
          2 : begin
              RB1:=1; C:=S;
              write(RB3,RB2,RB1)
            end
        end
      end
    until SYSTEM_BREAK
  end;

```

```

procedure CORE;
begin
  RESET_A;
  repeat
    NEXT_BYTE;
    if RI<=32 or RI=127
    then
      begin
        RESET_B;
        ISO646_CONYROL
      end
    else
      begin
        case C of
          0 : begin
              RB3:=RI; C:=1
            end;
          1 : if RI=33
              then CCCII_CONTROL
              else

```

```

        begin
            RB2:=RI; C:=2
        end
    2 : begin
        RB1:=1; C:=S;
        write(RB3,RB2,RB1)
    end
end
until SYSTEM_BREAK
end;

procedure CORTA;
begin
    case S of
        0 : write(RB3,RB2,RB1);
        1 : write(RB2,RB1);
        2 : write(RB1)
    end
end;

procedure OUT_LOCK_1(B3,B2);
begin
    case S of
        0 : begin
            write(33,33,115); { SSL code }
            write(B2,B3) { parameter }
        end;
        1 : begin
            write(33,115); { SSL code }
            write(B2,B3) { parameter }
        end
    end;
    S:=2
end;

procedure OUT_LOCK_2(B3);
begin
    case S of
        0 : begin
            write(33,33,112); { PSL code }
            write(34,B3); { parameter }
            S:=1
        end;
        1 : begin
            write(33,112); { PSL code }
            write(34,B3) { parameter }
        end
    end
end;
end;

```

```

procedure OUT_UNLOCK_1(B3,B2,B1)
begin
  if S=1
  then
    begin
      write(33,116); { SSU code }
      write(B2,B3,B1)
    end
  end;

procedure OUT_UNLOCK_2(B3,B2,B1);
begin
  case S of
    0 : begin
      write(33,33,114); { PSL code }
      write(34,B3,B2,B1) { parameter & 2-byte code }
    end;
    1 : begin
      write(33,114); { PSU code }
      write(34,B3,B2,B1) { parameter & 2-byte code }
    end
  end
end;

procedure OUT_LOCK_3;
begin
  case S of
    0 : write(33,119); { HOM code }
    1 : write(27,36,65) { ESC 2/4 4/1 }
  end
end;

procedure PSR_CODE;
begin
  write(33,113) { PSR code }
end;

```



1986

IFLA GENERAL CONFERENCE, TOKYO 1986

Division:

Section:

IFLA PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS, TOKYO

WHAT NEXT?

ISSUES ARISING FROM CONFERENCE DELIBERATIONS

by

Cynthia J. Durance  
Director  
Office for Network Development  
National Library of Canada  
Ottawa, Canada

Keywords:

automation of ideographic scripts  
automation of non-Roman alphabets  
character set standards  
transliteration

For internal use only:

To be brought to meeting no:

-----

Remarks:

「多言語・多文字資料利用のための図書館自動化システム—問題と解決」と題するIFLAプレコンファレンス・セミナーにおいて発表者が提示し、会議において認識された重要な問題を要約する。

ローマ文字を使用する言語のデータベースのための自動化方式が、必ずしも他の言語には適用できないことを、より一層認識する必要性が提起されている。このことは、表意文字(イデオグラフ)や非ローマ文字アルファベットのデータを処理する能力が、技術的進歩によって高められているとはいっても、国際的標準となる翻字法の欠如と、多くの言語を処理するためのコード化された文字セットの欠如が、自動化の大きな障壁となっているということである。IFLAは、これらの言語を自動化して行くための助言と支援を与え、そのための努力をおしまないだろう。そして、IFLAの目標である国際的書誌調整(UBC:ユニバーサル・ビブリオグラフィック・コントロール)の達成にむけて支援するものである。

検討された第一の論点は、ローマ字を使用したデータベースの自動化が、必ずしも他の言語の自動化に適応しないことの認識である。この問題を解決すべく、表意文字や非ローマ字のアルファベットが、主流を占めている国々の専門家達は、彼等にとって特有の必要性が、国際的により深く認識され、理解されて行くような道をさがさなくてはならない。

翻字は、いくつかの多言語データベースにとっては必要かつ有益なものであるが、最終的な解決法とは見なされていない。なぜならば、翻字データは、もとの言語を読めない人々にとってのみ有益なものだからである。そこで、さまざまな異なった翻字法と、国際的に認められた標準翻字法を確立する必要性がある。

多くの言語に適用しうる国際的に認められた標準文字セットの欠如もまた、自動化のためには大きな障壁と考えられている。これらの標準を取り決めて行くことは、重要かつ緊急の課題である。そのために、これらの言語が主流を占める国や地域の専門家の協力的な努力を必要としている。事実上の国際的標準として存在しているいくつかの標準規格は、主としてローマ字言語データの必要性に基づいて作られたものであるが、他の言語の必要性にも適用できるよう改正されなければならない。これらの標準規格なしには、自動化の経済的・社会的利益は完全には実現しないだろう。

パーソナル・コンピュータによって、複雑な言語や文字を入力し処理し表示する能力は大いに進歩してきた。将来、エキスパート・システム技術やデータ辞書システムといった新しい開発成果は、複雑なデータを整理し調整するための、もっと柔軟な手段を、広範に提供してくれるであろう。

IFLAの目標とする普遍的な書誌調整を達成するためには、このプレコンファレンスのすべての発表者がなんらかの形で指摘したように、自動化の開発は不可欠のものである。IFLAは、その核となる様々な計画を通じて、この大きな仕事を多方面から援助する助力をおしまないだろう。そしてまた、出版、セミナー、地域的話し合いなどの場を通じて、IFLAはすべての言語と文字のデータベースを構築するという大事業に対して、より深い理解を示し、これを推進していくものである。

# ABSTRACT

The paper provides a summary of the major issues identified by speakers at the IFLA pre-conference seminar on "Automated Systems for Access to Multilingual and Multiscript Library Materials. These include the need for increased awareness that automated solutions for Roman-language data bases are not always applicable to other languages. It is recognized that lack of international standard transliteration schemes and coded character sets for many languages are major barriers to their automation, even though technological advances are increasing the ability to manipulate ideographic and non-Roman alphabetic data. IFLA can advise on and assist efforts to automate these languages and thereby help achieve its goal of Universal Bibliographic Control.



The task I have been given, to summarize the deliberations of the past two days, is a daunting one. Although I have considerable experience with automated systems and bilingual bibliographic control in Roman languages, unlike all of the other speakers at this pre-conference seminar, I have no experience in the automation of languages in other alphabets or of ideographic scripts.

I have learned a great deal by reading and analyzing the papers given during this seminar. My overall impression is one of great admiration for those who are working so creatively to solve the immensely difficult problems of automating languages that are written in ideograms or alphabets other than Roman.

In the early 1970's, we at the National Library of Canada developed the first bilingual MARC format, and like UTLAS, undertook to create a bilingual data base in English and French. At that time we thought it a very difficult and challenging task - and it was. Now, after this seminar, I realize we didn't know the meaning of the word "difficult". Solving the problems of handling not only different languages but different scripts and syntaxes and also linguistic complexities such as contextual and phonetic as well as written language, is many orders of magnitude more difficult than processing different Roman languages.

I would like to summarize those areas our speakers have identified as requiring action in order that we can further the automation of ideographic, non-Roman alphabet, and multiscript data bases.

The first issues are issues of awareness and influence. In the western world there is perhaps an insufficient appreciation that the technology

and standard practices which permit automated bibliographic control for Roman languages are not necessarily applicable to other languages because they cannot handle the full range of problems encountered.

On the other hand, those countries which require ideographic and non-Roman alphabet language processing have perhaps not been as vigorous and persistent as they could have been in making their needs known to international bodies. I think it is fair to say that the International Organization for Standardization (ISO) and the International Federation of Library Associations (IFLA) have been dominated by representatives from the western world. This is not to criticize these bodies, but rather to point out that they are volunteer organizations, with the result that those countries who are prepared to participate actively in their deliberations will be the ones who ultimately have the greatest influence. The challenge is to find ways in which other countries, such as the Asian and Arabic countries, can more fully participate in and influence the work of these organizations. These issues of awareness and influence are difficult problems, not easily resolved, compounded as they are by the barriers of language, culture, politics, distance, and cost, but perhaps IFLA seminars like this one are a step in the right direction.

The second issue is the issue of transliteration. Transliteration has its uses for certain multi-language data bases and for certain functions. The International Serials Data System (ISDS) data base is an example. The purpose of that data base is to identify serial publications published world-wide and currently data in 113 languages are being processed. In such a program there seems to be no other solution in the near term but to create the data base in a single alphabet.

In situations where transliteration seems the best solution, and here I speak either of transliteration to the Roman alphabet or to another alphabet or script, we suffer from too many non-standard schemes. The current multiple transliteration schemes in use from Chinese, Japanese and Korean to Roman can only serve to complicate efforts to achieve universal bibliographic control for these languages. Add to these schemes the multiple conventions for transliterating Kana to Hangul or vice versa, not to mention to other alphabets, and the problem is compounded exponentially. I think it hardly needs to be stated that we must achieve and use internationally recognized standard transliteration schemes. I have seen a work item of the ISO Technical Committee 46 for a transliteration standard from Korean to Roman (see appendix 1). This is a very encouraging development.

Although transliteration must sometimes be used in performing library and information functions, its shortcomings are so severe that such techniques cannot be considered an optimal solution to the processing of bibliographic and information data bases. As we all know, transliteration is only useful for librarians who do not read the original language. It is confusing, if not completely incomprehensible, to both librarians and users who do read the original language. Advances in technology are now beginning to provide some alternatives to transliterated data, and we have heard here of some exciting attempts to use technology to provide automated data bases designed for those who are conversant with non-Roman alphabet and ideographic languages.

The third and most pressing issue seems to be that of character set standards. Currently, there are approved ISO character set standards for African, Cyrillic, Greek and Latin (Roman) languages. In addition Arabic, Armenian, Georgian and Hebrew are in various stages of the ISO

standardization process (see Appendix 1). Hopefully, these emerging standards will take into account the requirements of related languages as well.

Although this list of accomplishments is quite impressive, there are major omissions. A good deal of work has been done within national boundaries to develop character sets for Chinese, Japanese and Korean, but there is an urgent need to agree internationally on a single basic character set for written Chinese that will also handle written Japanese and Korean. An international and recognized mechanism for registering extensions as well as historical and cultural variations to that Chinese character set is also required. A character set standard for Thai is also not yet in the ISO work plan.

Yet another series of standards which ISO needs to be asked to consider are phonetic character sets for Japanese and Korean as these are often used for arranging and sorting data. It is my understanding that there are national or de facto national standards for these, and if so, their approval by ISO could well be a rather simple process.

Although there are arbitrary techniques for filing ideographic languages which can be incorporated into automated processes, such as the system of filing Chinese characters based on an analysis of the radicals and strokes, this does not solve completely the problem of effective automated information retrieval. The extent to which widely used conventions can be adopted for effective information retrieval based on phonetic character sets or other means will affect the ease with which users can search both single and multiple data bases. These conventions take on even greater importance in multiscrypt data bases. I believe it is fair to say that, to date, more effort has been expended on the



problems of input and display of non-Roman alphabet and ideographic character sets than has been devoted to the attendant and inter-related problems of information retrieval from these data bases. This is an area which requires more attention.

Successful completion of any of this large list of complex undertakings will require the concerted effort, time, and negotiating skill of both language experts and computer experts not only in each country, but also amongst experts of those countries which have common language roots, and finally approval by ISO. If countries which have common language roots can agree amongst themselves on a character set, they will greatly increase the chances of their recommendations being accepted by ISO. Reaching agreement on standards is a time-consuming, frustrating and difficult task, but unless a common way can be found to represent and encode data, the social and economic benefits of automated bibliographic control nationally or internationally certainly will be compromised, if not actually jeopardized.

More than one speaker has noted that experts in the western world, although their expertise can be of major assistance, should alone not make the decisions on character sets for non-Roman languages. If they do there is a risk of the resulting standards not fully meeting the needs of the principal users of the standard. It is in countries where a language is predominant that dependence on the standard is greatest: therefore the decisions on how these languages are encoded should above all reflect their perception of their needs.

Some existing de facto international standards which were developed based primarily on the requirements of Roman languages may need to be revised or adjusted to more fully take into account the requirements of non-Roman languages. One example is the set of International Standard Bibliographic Descriptions (ISBD). For instance, I am told that the long dash and period used in the ISBD's to separate data elements is almost identical to the symbol for the numeral "1" in written Chinese. Another standard which may need change is UNIMARC where extensions and/or adjustments are required for ideographic language data. In this connection I believe it is important to emphasize that all of us who adopt standards or de facto standards have a real and on-going responsibility to ensure that any changes or additions we require are formally negotiated with and recognized by the responsible international body. It is not enough to say we follow a standard, we must all have a formal and permanent national mechanism for submitting requested changes and enhancements. Otherwise, divergence becomes greater and greater until the standard is a standard in name only. Adjustments to de facto standards such as the ISBD's or UNIMARC, if necessary, may not always be easily negotiated because of the widespread incorporation of them in existing library systems in the western world, and the resulting cost of change.

One of our speakers mentioned the difficulty of persuading hardware and software manufacturers to implement systems for coding languages which are not widely used. Achievement of recognized standard character sets for those languages is an incentive for implementation. Even if the market remains small, financial assistance from international or other organizations is more easily acquired if the request is to implement a recognized standard, rather than a non-standard character set.



The fourth issue is that of the technology itself. In the past twenty years the state of the art of hardware and software has advanced greatly, aided in certain instances by the technological innovations required to display ideographic languages. A case in point is the dot matrix printer which was invented in Asia, I believe in Japan, to print ideograms, and which is now providing elegant and eye-pleasing computer printouts worldwide. The widespread use of the personal computer in information technology has made possible major advances in the ability to input, output, display and edit complex alphabets and scripts. All indications are that these advances in the capacity and functionality of affordable workstations will continue and will provide increasingly flexible ways to process data.

As expert systems (also called artificial intelligence) techniques become more widely used in computer programming we likely will see further advances in the ability to convert and manipulate complex data. Data dictionary based systems, now increasingly widely used, may prove to be the solution to the problem of registering and disseminating information about extensions to ideographic character sets and of accommodating historical and cultural variations to these character sets. International standards for telecommunications now being widely implemented will ease long distance transmission and allow easier access to remote systems. Continuing advances in technology will provide ever greater possibilities for processing complex data, but if we wish to be able to take full advantage of them we must continue to position ourselves to do so. I believe our challenge is not so much the technology, but the lack of data standards which frustrates our ability to exploit the technology we have available to us.

The use of automated developments to process data in non-Roman alphabets and scripts, such as our speakers have examined, is essential if we are serious about achieving the IFLA concept of Universal Bibliographic Control. However, much remains to be done before we will have bibliographic control in any true universal sense. We have a saying in English which I think applies to this goal IFLA has set for itself. It is "the difficult we do immediately: the impossible we do, but it just takes a little longer".

What can IFLA do to help? What can we do to help IFLA in this monumental task?

The UBC office can certainly offer assistance and information on existing standards for bibliographic data, and can provide information about experts in bibliographic control, and character sets in other countries.

The National Library of Canada as host institution for the IFLA International Office for Transborder Data Flow will in the coming year be providing a clearinghouse of information for IFLA members on new developments in library automation and in information technology.

The International Marc Programme at the Deutschebibliothek in Germany maintains the UNIMARC format and provides information about Marc formats for IFLA members. They would certainly be pleased to discuss any problems and concerns.

IFLA can support you and encourage you and it may, if requested, be able to assist in finding sources of funding for some tasks. However, we must remember that IFLA does not itself have great financial resources to draw upon.

IFLA can also, through publications, seminars like this one, and by sponsoring international or regional discussion groups on major issues, serve to focus attention on and raise the level of understanding of the challenging problems of processing data in non-Roman alphabets and ideographic scripts.

But we must all remember, as we talk of IFLA as an organization, that we are IFLA - the organization only exists because we, the librarians of the world, support it, and only by working together to propose and realize innovative programs can we make it a dynamic organization that meets our needs. This is our responsibility.

In closing, I would like to quote from the first paper given at this seminar, and thereby close the circle of our deliberations. Dr. Nicolescu said,

"I believe that the welfare of an individual nation should and can coincide with that of regional and global communities. I also believe that this statement applies as well to information systems and services".<sup>1</sup>

I profoundly echo her belief. It seems an appropriate note on which to close this, the last of the formal papers of this seminar.

---

1) Har-Nicolescu, Suzine. "Romanized and transliterated data bases of Asian-language materials: history, problems and prospects." Paper presented at the IFLA pre-conference seminar on Automated Systems for Access to Multilingual and Multiscript Library Materials. Tokyo, August, 1986. p. 16.

Status of Work related to character sets  
within ISO/TC 46 and ISO TC/97\*

---

| TITLE                                                                                                                   | STATUS                |
|-------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Arabic coded character set                                                                                              | 1st working draft     |
| Documentation - Hebrew alphabet character sets for bibliographic information interchange                                | DP 8957 <sup>1</sup>  |
| Documentation - Cyrillic alphabet coded character sets for slavonic languages for bibliographic information interchange | DIS 6861 <sup>2</sup> |
| Documentation - Mathematical coded character set for bibliographic information interchange                              | DIS 6862              |
| Documentation - Bibliographic control characters                                                                        | DIS 6630              |
| Extension of the Latin alphabet coded character set for bibliographic information interchange                           | ISO 5426 <sup>3</sup> |
| Documentation - African coded character set for bibliographic information interchange                                   | ISO 6438              |
| Extension of the Cyrillic alphabet coded character set for bibliographic information interchange                        | ISO 5427              |
| Greek alphabet coded character set for bibliographic information interchange                                            | ISO 5428              |
| Phonetic coded character set for bibliographic information interchange                                                  | NWI <sup>4</sup>      |
| Transliteration of Korean characters into Latin Characters                                                              | NWI                   |

---

\* This list does not presume to be comprehensive

- 1) DP - draft proposal
- 2) DIS - draft international standard
- 3) ISO----- - standard
- 4) NWI - new work item

|                                                                                                                                                               |                     |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Coded character set for Latin characters used in minor European languages and in obsolete typography for bibliographic information interchange                | NWI                 |
| Armenian coded character set for bibliographic information interchange                                                                                        | NWI                 |
| Georgian coded character set for bibliographic information interchange                                                                                        | NWI                 |
| Handling of character sets and typographical symbols in electronic publishing                                                                                 | NWI, soon to be DIS |
| Information processing - Implementation of the 7-bit coded character set and its 7-bit and 8-bit extensions on 3,81 mm magnetic cassette for data interchange | ISO 3275            |
| Information processing - 8-bit coded character set for information interchange                                                                                | ISO 4873            |
| Information processing - ISO 7-bit and 8-bit coded character sets. Additional control functions for character-imaging devices                                 | ISO 6429            |
| Data processing - Implementation of the ISO 7-bit and 8-bit coded character sets on punched card                                                              | ISO 6586            |
| Data processing - Conversion between the ISO 7-bit coded character set (ISO 646) and the CCITT international telegraph alphabet No. 2 (ITA 2)                 | ISO 6936            |
| Information processing - Coded character sets for text communication - Part 1: General introduction                                                           | ISO 6937/1          |
| Information processing - Coded character sets for text communication - Part II: Latin alphabetic and non-alphabetic graphic characters                        | ISO 6937/2          |
| Information processing - ISO 7-bit coded character set for information interchange                                                                            | ISO 646             |
| Information processing - Implementation of the 7-bit coded character set and its 7-bit and 8-bit extensions on 9-track 12,7 mm (0.5 in) magnetic tape         | ISO 962             |
| Information processing - Guide for the definition of 4-bit character sets derived from the 7-bit coded character set for information processing interchange   | ISO 963             |



|                                                                                                                                    |            |
|------------------------------------------------------------------------------------------------------------------------------------|------------|
| Alphanumeric character sets for optical recognition<br>- Part I: Character set OCR-A - Shapes and dimensions of the printed image  | ISO 1073/1 |
| Alphanumeric character sets for optical recognition<br>- Part II: Character set OCR-B - Shapes and dimensions of the printed image | ISO 1073/2 |
| Information processing - Representation of the 7-bit coded character set on punched tape                                           | ISO 1113   |
| Information processing - Character structure for start/stop and synchronous character oriented transmission                        | ISO 1177   |
| Hardware representation of ALGOL basic symbols in the ISO 7-bit coded character set for information processing interchange         | ISO 1672   |
| Printing specifications for optical character recognition                                                                          | ISO 1831   |
| Information processing - ISO 7-bit and 8-bit coded character sets - Code extension techniques                                      | ISO 2022   |
| Information processing - Coding of machine readable characters (MICR ANS OCR)                                                      | ISO 2033   |
| Information processing - Graphical representations for the control characters of the 7-bit coded character set                     | ISO 2047   |
| Keyboard for international information processing interchange using the ISO 7-bit coded character set<br>-Alphanumeric area        | ISO 2530   |



IFLA

IFLA GENERAL CONFERENCE, TOKYO 1986

Division:

Section:

1986

IFLA PRE-CONFERENCE SEMINAR ON AUTOMATED SYSTEMS  
FOR ACCESS TO MULTILINGUAL AND MULTISCRIP  
LIBRARY MATERIALS, TOKYO

168-MCULTP-11-E  
INF

WHAT NEXT?

ISSUES ARISING FROM CONFERENCE DELIBERATIONS

by

Cynthia J. Durance  
Director  
Office for Network Development  
National Library of Canada  
Ottawa, Canada

Keywords:

automation of ideographic scripts  
automation of non-Roman alphabets  
character set standards  
transliteration

Internal use only:

be brought to meeting no:

Friday 22 August

marks:

be presented orally X

BEST COPY AVAILABLE

# ABSTRACT

The paper provides a summary of the major issues identified by speakers at the IFLA pre-conference seminar on "Automated Systems for Access to Multilingual and Multiscript Library Materials. These include the need for increased awareness that automated solutions for Roman-language data bases are not always applicable to other languages. It is recognized that lack of international standard transliteration schemes and coded character sets for many languages are major barriers to their automation, even though technological advances are increasing the ability to manipulate ideographic and non-Roman alphabetic data. IFLA can advise on and assist efforts to automate these languages and thereby help achieve its goal of Universal Bibliographic Control.

The task I have been given, to summarize the deliberations of the past two days, is a daunting one. Although I have considerable experience with automated systems and bilingual bibliographic control in Roman languages, unlike all of the other speakers at this pre-conference seminar, I have no experience in the automation of languages in other alphabets or of ideographic scripts.

I have learned a great deal by reading and analyzing the papers given during this seminar. My overall impression is one of great admiration for those who are working so creatively to solve the immensely difficult problems of automating languages that are written in ideograms or alphabets other than Roman.

In the early 1970's, we at the National Library of Canada developed the first bilingual MARC format, and like UTLAS, undertook to create a bilingual data base in English and French. At that time we thought it a very difficult and challenging task - and it was. Now, after this seminar, I realize we didn't know the meaning of the word "difficult". Solving the problems of handling not only different languages but different scripts and syntaxes and also linguistic complexities such as contextual and phonetic as well as written language, is many orders of magnitude more difficult than processing different Roman languages.

I would like to summarize those areas our speakers have identified as requiring action in order that we can further the automation of ideographic, non-Roman alphabet, and multiscript data bases.

The first issues are issues of awareness and influence. In the western world there is perhaps an insufficient appreciation that the technology

and standard practices which permit automated bibliographic control for Roman languages are not necessarily applicable to other languages because they cannot handle the full range of problems encountered.

On the other hand, those countries which require ideographic and non-Roman alphabet language processing have perhaps not been as vigorous and persistent as they could have been in making their needs known to international bodies. I think it is fair to say that the International Organization for Standardization (ISO) and the International Federation of Library Associations (IFLA) have been dominated by representatives from the western world. This is not to criticize these bodies, but rather to point out that they are volunteer organizations, with the result that those countries who are prepared to participate actively in their deliberations will be the ones who ultimately have the greatest influence. The challenge is to find ways in which other countries, such as the Asian and Arabic countries, can more fully participate in and influence the work of these organizations. These issues of awareness and influence are difficult problems, not easily resolved, compounded as they are by the barriers of language, culture, politics, distance, and cost, but perhaps IFLA seminars like this one are a step in the right direction.

The second issue is the issue of transliteration. Transliteration has its uses for certain multi-language data bases and for certain functions. The International Serials Data System (ISDS) data base is an example. The purpose of that data base is to identify serial publications published world-wide and currently data in 113 languages are being processed. In such a program there seems to be no other solution in the near term but to create the data base in a single alphabet.

In situations where transliteration seems the best solution, and here I speak either of transliteration to the Roman alphabet or to another alphabet or script, we suffer from too many non-standard schemes. The current multiple transliteration schemes in use from Chinese, Japanese and Korean to Roman can only serve to complicate efforts to achieve universal bibliographic control for these languages. Add to these schemes the multiple conventions for transliterating Kana to Hangul or vice versa, not to mention to other alphabets, and the problem is compounded exponentially. I think it hardly needs to be stated that we must achieve and use internationally recognized standard transliteration schemes. I have seen a work item of the ISO Technical Committee 46 for a transliteration standard from Korean to Roman (see appendix 1). This is a very encouraging development.

Although transliteration must sometimes be used in performing library and information functions, its shortcomings are so severe that such techniques cannot be considered an optimal solution to the processing of bibliographic and information data bases. As we all know, transliteration is only useful for librarians who do not read the original language. It is confusing, if not completely incomprehensible, to both librarians and users who do read the original language. Advances in technology are now beginning to provide some alternatives to transliterated data, and we have heard here of some exciting attempts to use technology to provide automated data bases designed for those who are conversant with non-Roman alphabet and ideographic languages.

The third and most pressing issue seems to be that of character set standards. Currently, there are approved ISO character set standards for African, Cyrillic, Greek and Latin (Roman) languages. In addition Arabic, Armenian, Georgian and Hebrew are in various stages of the ISO



standardization process (see Appendix 1). Hopefully, these emerging standards will take into account the requirements of related languages as well.

Although this list of accomplishments is quite impressive, there are major omissions. A good deal of work has been done within national boundaries to develop character sets for Chinese, Japanese and Korean, but there is an urgent need to agree internationally on a single basic character set for written Chinese that will also handle written Japanese and Korean. An international and recognized mechanism for registering extensions as well as historical and cultural variations to that Chinese character set is also required. A character set standard for Thai is also not yet in the ISO work plan.

Yet another series of standards which ISO needs to be asked to consider are phonetic character sets for Japanese and Korean as these are often used for arranging and sorting data. It is my understanding that there are national or de facto national standards for these, and if so, their approval by ISO could well be a rather simple process.

Although there are arbitrary techniques for filing ideographic languages which can be incorporated into automated processes, such as the system of filing Chinese characters based on an analysis of the radicals and strokes, this does not solve completely the problem of effective automated information retrieval. The extent to which widely used conventions can be adopted for effective information retrieval based on phonetic character sets or other means will affect the ease with which users can search both single and multiple data bases. These conventions take on even greater importance in multiscript data bases. I believe it is fair to say that, to date, more effort has been expended on the



problems of input and display of non-Roman alphabet and ideographic character sets than has been devoted to the attendant and inter-related problems of information retrieval from these data bases. This is an area which requires more attention.

Successful completion of any of this large list of complex undertakings will require the concerted effort, time, and negotiating skill of both language experts and computer experts not only in each country, but also amongst experts of those countries which have common language roots, and finally approval by ISO. If countries which have common language roots can agree amongst themselves on a character set, they will greatly increase the chances of their recommendations being accepted by ISO. Reaching agreement on standards is a time-consuming, frustrating and difficult task, but unless a common way can be found to represent and encode data, the social and economic benefits of automated bibliographic control nationally or internationally certainly will be compromised, if not actually jeopardized.

More than one speaker has noted that experts in the western world, although their expertise can be of major assistance, should alone not make the decisions on character sets for non-Roman languages. If they do there is a risk of the resulting standards not fully meeting the needs of the principal users of the standard. It is in countries where a language is predominant that dependence on the standard is greatest: therefore the decisions on how these languages are encoded should above all reflect their perception of their needs.

Some existing de facto international standards which were developed based primarily on the requirements of Roman languages may need to be revised or adjusted to more fully take into account the requirements of non-Roman languages. One example is the set of International Standard Bibliographic Descriptions (ISBD). For instance, I am told that the long dash and period used in the ISBD's to separate data elements is almost identical to the symbol for the numeral "1" in written Chinese. Another standard which may need change is UNIMARC where extensions and/or adjustments are required for ideographic language data. In this connection I believe it is important to emphasize that all of us who adopt standards or de facto standards have a real and on-going responsibility to ensure that any changes or additions we require are formally negotiated with and recognized by the responsible international body. It is not enough to say we follow a standard, we must all have a formal and permanent national mechanism for submitting requested changes and enhancements. Otherwise, divergence becomes greater and greater until the standard is a standard in name only. Adjustments to de facto standards such as the ISBD's or UNIMARC, if necessary, may not always be easily negotiated because of the widespread incorporation of them in existing library systems in the western world, and the resulting cost of change.

One of our speakers mentioned the difficulty of persuading hardware and software manufacturers to implement systems for coding languages which are not widely used. Achievement of recognized standard character sets for those languages is an incentive for implementation. Even if the market remains small, financial assistance from international or other organizations is more easily acquired if the request is to implement a recognized standard, rather than a non-standard character set.

The fourth issue is that of the technology itself. In the past twenty years the state of the art of hardware and software has advanced greatly, aided in certain instances by the technological innovations required to display ideographic languages. A case in point is the dot matrix printer which was invented in Asia, I believe in Japan, to print ideograms, and which is now providing elegant and eye-pleasing computer printouts worldwide. The widespread use of the personal computer in information technology has made possible major advances in the ability to input, output, display and edit complex alphabets and scripts. All indications are that these advances in the capacity and functionality of affordable workstations will continue and will provide increasingly flexible ways to process data.

As expert systems (also called artificial intelligence) techniques become more widely used in computer programming we likely will see further advances in the ability to convert and manipulate complex data. Data dictionary based systems, now increasingly widely used, may prove to be the solution to the problem of registering and disseminating information about extensions to ideographic character sets and of accommodating historical and cultural variations to these character sets. International standards for telecommunications now being widely implemented will ease long distance transmission and allow easier access to remote systems. Continuing advances in technology will provide ever greater possibilities for processing complex data, but if we wish to be able to take full advantage of them we must continue to position ourselves to do so. I believe our challenge is not so much the technology, but the lack of data standards which frustrates our ability to exploit the technology we have available to us.

The use of automated developments to process data in non-Roman alphabets and scripts, such as our speakers have examined, is essential if we are serious about achieving the IFLA concept of Universal Bibliographic Control. However, much remains to be done before we will have bibliographic control in any true universal sense. We have a saying in English which I think applies to this goal IFLA has set for itself. It is "the difficult we do immediately: the impossible we do, but it just takes a little longer".

What can IFLA do to help? What can we do to help IFLA in this monumental task?

The UBC office can certainly offer assistance and information on existing standards for bibliographic data, and can provide information about experts in bibliographic control, and character sets in other countries.

The National Library of Canada as host institution for the IFLA International Office for Transborder Data Flow will in the coming year be providing a clearinghouse of information for IFLA members on new developments in library automation and in information technology.

The International Marc Programme at the Deutschebibliothek in Germany maintains the UNIMARC format and provides information about Marc formats for IFLA members. They would certainly be pleased to discuss any problems and concerns.

IFLA can support you and encourage you and it may, if requested, be able to assist in finding sources of funding for some tasks. However, we must remember that IFLA does not itself have great financial resources to draw upon.

IFLA can also, through publications, seminars like this one, and by sponsoring international or regional discussion groups on major issues, serve to focus attention on and raise the level of understanding of the challenging problems of processing data in non-Roman alphabets and ideographic scripts.

But we must all remember, as we talk of IFLA as an organization, that we are IFLA - the organization only exists because we, the librarians of the world, support it, and only by working together to propose and realize innovative programs can we make it a dynamic organization that meets our needs. This is our responsibility.

In closing, I would like to quote from the first paper given at this seminar, and thereby close the circle of our deliberations. Dr. Nicolescu said,

"I believe that the welfare of an individual nation should and can coincide with that of regional and global communities. I also believe that this statement applies as well to information systems and services".<sup>1</sup>

I profoundly echo her belief. It seems an appropriate note on which to close this, the last of the formal papers of this seminar.

---

1) Har-Nicolescu, Suzine. "Romanized and tranliterated data bases of Asian-language materials: history, problems and prospects." Paper presented at the IFLA pre-conference seminar on Automated Systems for Access to Multilingual and Multiscript Library Materials. Tokyo, August, 1986. p. 16.



The review of the International Standard Bibliographic Descriptions (ISBDs)  
and non-roman script publications

by Barbara Jover, UBC Programme Officer, IFLA International  
Programme for UBC, c/o The British Library, Great Russell  
Street, LONDON, UK

When in 1980 the Standing Committee of the Section on Cataloguing agreed to review the texts of the ISBDs it was stated that one of the tasks of the Review Group would be to "consider the particular problems of non-roman scripts, and especially those written right-to-left, in order to ensure that all the ISBDs are equally hospitable to publications in all types of scripts". Consequently, the Review Committee requested and received a number of comments and suggestions from cataloguers working with these materials.

The comments were carefully considered and the following discussion, taken from the Review Group papers, considers in general terms the following areas where requests for changes were made: ISBD punctuation, printing terminology and use of colophon. Following this are extracts from the draft of the second edition of ISBD(M) dealing specifically with non-roman scripts.

- 1 The charge is often made that ISBD punctuation is alien to most non-roman languages and that the vernacular equivalences of the ISBD marks should be allowed. Although much of the European/North American library world has been applying ISBD punctuation for nearly ten years, many US librarians first argued that using slashes, equals signs, etc. in bibliographic records would confuse library patrons and stressed their unnatural appearance, as the English-language punctuation system did not use many of the ISBD marks, at least not with the ISBD significance. Furthermore, there was no *natural* English mark in existence to express such concepts as statement of responsibility or parallel title. The same thing could be said of any other roman language. Therefore there seems to be no real difference between roman and non-roman languages in this respect, except for the obvious fact that the marks used by ISBD all originate in roman linguistics, mathematics, etc. Just as librarians, dealing with roman script material have accepted the more or less alien ISBD marks, so those handling non-roman scripts could do likewise. In addition, one of the primary values of the uniform punctuation is to help users to identify the elements of the entry when they are not



proficient in the language. The uniform punctuation becomes in effect a universal sign language.

- 2 Another frequently voiced comment relates to the fact that the books in non-roman languages do not normally show words that distinguish between the concepts of editing and publishing, or between the concepts of printing and publishing. This is not a phenomenon peculiar to non-roman publications. Many roman languages show the same ambiguity, e.g. Spanish, French, and Hungarian. This is not something that either a standard or rules can deal with. In these cases one must depend upon the skill and experience of the cataloguer in making the necessary distinctions when viewing ambiguous phrases on publications. At any rate, the major aspect of the problem is the difficulty of distinguishing between publishing statements and impression (printing) statements. Here it may be necessary to have rather arbitrary guidelines in certain areas, but these must be left to national cataloguing codes.
- 3 Several comments related to the presence of the colophon in many non-roman language publications, pointing out that at least in some instances the colophon, rather than the title page, should be considered the chief source of information. In Asian books particularly the colophon information is often much fuller than one would find in roman-language colophons. Tradition, or perhaps even legal requirements past or present, require that in these cases full, authoritative data appear in the colophon, while the title page is left to the imaginative flights of the publisher or author. Some publications from Japan and Taiwan are the best example of this phenomenon. At least in modern books, however, a marked adoption of the Western-style title page is shown. It is not that the colophon is disappearing. Rather, the title page has changed, showing more often than not approximately the same amount and kind of data an average Western book would show. Publications from mainland China and Korea afford numerous examples. Thus it would be going against the rather clear trend of Asian publications to promote the colophon in a way that would demote the title page.

Consultation with specialists at the Library of Congress, however, has shown that approximately 30% of current Japanese publications do not follow this trend. In these cases, rarely do the following elements appear on the title page: other title information, statement of responsibility, edition statement, place of publication, and publication date. While the percentage for Chinese and Korean is not as high, there is a significant number of comparable examples. Thus for Chinese, Japanese, and Korean, a combination of the title page and colophon should be treated as the single chief source, thus promoting the colophon without demoting the title page.

- 4 A few comments suggested that the data in areas 5, 7, and 8 (physical description, notes, and ISBNs and price) should be stated in the language of the publications rather than in that of the cataloguing agency. At least physical description and notes have been traditionally given in the language of the cataloguing agency and it seems most unlikely that a change could be contemplated. Again, the need, if it exists, is not peculiar to non-roman publications, since the many European languages collected in libraries worldwide show the same phenomena and the same consequent possibilities. At the same time, one can appreciate the concern of some non-roman cataloguers for the time and trouble it causes them to shift continually from one typewriter to another. This would be true not only in the cases of recording areas 1-4 in one language and areas 5, 7, and 8 in another, but also in cases of parallel elements in areas 1 or 2. It seems, however, that this problem must be solved by technology, rather than by the ISBD recommendations.
- 5 The remaining points gleaned from comments that did not become recommendations are details of bibliographic description. It was agreed that the complexities of pagination or foliation are more properly dealt with in national cataloguing codes than in the ISBDs.

The following are extracts from *ISBD(M) second edition*, illustrating the changes made to accommodate publications in non-roman scripts:

#### Punctuation

0.4.11 When information is given in scripts written from right to left, commas and semi-colons used as prescribed punctuation are reversed when that is the style of the script. Similarly, the point, space, dash, space combination of prescribed punctuation reads from right to left and the meanings of open and closed parentheses and square brackets are reversed. The diagonal slash and groups of western Arabic numerals which are not reversed in such scripts are not reversed when given.

#### 0.5 Sources of information

When a publication lacks a title-page, an alternate source of information is selected as the title-page substitute. The selection of the source to be treated as title-page substitute is determined by

consideration of which source has the fullest information (e.g., the colophon might be preferred over any of the preliminaries).

In the case of oriental publications in non-roman scripts where full bibliographic details are given in the colophon, a leaf standing in the position of a title-page and bearing the title proper is not to be considered the title-page in the following circumstances:

- A. when the leaf bears only the title proper in the manner of a half-title page;
- B. when the leaf bears the title proper, with or without other bibliographic information, in a calligraphic version (the full bibliographic details in the colophon being given in conventional forms of Chinese characters as used in modern Chinese, Japanese and Korean printing);
- C. when the leaf bears only a western language version of the title and/or publication details.

In each of these cases the first preference for the title-page substitute is the colophon.

#### 0.6 Language and script of the description

Elements in areas 1, 2, 4 and 6 are normally transcribed from the publication and are, therefore, wherever practicable, in the language and/or script in which they appear there. Interpolations in these areas are enclosed in square brackets and are given in the language and/or script of the context, except:

- prescribed abbreviations (see 0.7) and prescribed interpolations (see 0.10);
- general material designation (see 1.2) and statement of function of distributor (see 4.3) which, when supplied, are given in the language and/or script chosen by the national bibliographic agency.

Terms used in areas 5, 7 and 8 are not enclosed in square brackets and are given in the language and/or script chosen by the national bibliographic agency, except:

- when original title or variant title is provided in area 7;
- when quotations are provided in area 7.

The description of publications appearing in scripts other than that used by the national bibliographic agency may, if necessary, be transliterated or transcribed without brackets into the script used by the agency.

## 0.7 Abridgements and abbreviations

- 0.7.1 In exceptional cases the abridgement of certain elements in the description is permitted, provided the omission takes place at the end or in the middle of the element (e.g. a lengthy title proper, see 1.1.4.1). In such cases, the omission is indicated by marks of omission.
- 0.7.2 In the case of the abridgement of a single statement of responsibility consisting of the names of several persons or corporate bodies (see 1.5.4.3), the omission is indicated by marks of omission and the insertion of the prescribed abbreviation "et al". (= et alii, and others) enclosed in square brackets (or for non-roman script records, its equivalent in another script).

## 0.8 Capitalization

In general, the first letter of the first word of each area and some elements (e.g. parallel title, alternative title, or section title) should be a capital. The first letter of the general material designations should also be a capital. Other capitalization should follow the appropriate usage for the language(s) and/or script(s) used in the description (see 0.6). When more than one language and/or script appears in the description, each should be capitalized in accordance with the usage of that language and/or script even when this produces an inconsistent pattern of capitalization for the description as a whole.

## Appendix B

### Bi-directional records

Bi-directional records are multi-script records in scripts written from left to right and scripts written from right to left. Unless the direction of the script changes within an element each change of script begins on a new line of the description at the appropriate margin. Prescribed punctuation is given at the start of the new element in the style required by the script, except for the prescribed point, comma or semi-colon which is given at the end of the previous element in the style required by the script (see also 0.4 Punctuation). Thus, the point of the point, space, dash, space (. -) ends the previous element and the dash begins the following element on a new line at the appropriate margin.

#### ISBD(S) examples:

Examples with Jawi script which reads from right to left:

##### Example 1

سبر جمعہ . — تاہن 1، بیل 1، (سپتمبر 1980) -  
 . — کوالا لپور : بہاگین اکام ، جاتن فردانا منتری ،  
 1980 - . — جیلید : 31 سینتی میٹر .  
 دوسلن .  
 فرچوما

Example 1 shows the ISBD mirror image punctuations that are used in the Jawi script (written from right-to-left) i.e. ، (for comma), ؛ (semi-colon). This example also illustrates the punctuation pattern for Jawi script for the punctuation point, space, dash, space. Each new area is preceded by a space, dash, space, point (— .)

##### Example 2

Ar-Ridzwan

الرضوان / جاتن حال احوال اکام فیرق .  
 — Bil. 1 (Dis. 1978) - . — Ipoh :  
 Jabatan Hal Ehwal Agama Perak, 1978- .  
 — جیلید : ایلوستراسی، 27 سینتی میٹر

Example 2 shows a record with the title proper in the Roman script (left-to-right) and parallel title and author statement in the Jawi script (right-to-left). The imprint is in the Roman script.



Bi-directional records

Bi-directional records are multi-script records in scripts written from left to right and scripts written from right to left.

Unless the direction of the script changes within an element each change of script begins on a new line of the description at the appropriate margin. Prescribed punctuation is given at the start of the new element in the style required by the script, except for the prescribed point of comma which is given at the end of the previous element in the style required by the script (see also 0.4 Punctuation). Thus the point of the point, space, dash, space ( . - ) ends the previous element and the dash begins the following element on a new line at the appropriate margin.

Examples : (ISBD(M) examples)

القاموس الحديث، فرنسي - عربي / تأليف ميري إلياس

= Dictionnaire moderne, français - arabe / par Mitri Elias.

- طبعة 4.4، منقحة ومزودة

= 4me éd., rev. et augm.

- القاهرة : الطبعة المعاصرة

= Caire : Elias' Modern Press, 1974.

(Cataloged with Arabic data transcribed first)

Dictionnaire moderne, français - arabe / par Mitri Elias

= القاموس الحديث، فرنسي - عربي / تأليف ميري إلياس.

- 4me éd., rev. et augm.

= طبعة 4.4، منقحة ومزودة.

- Caire : Elias' Modern Press

= القاهرة : الطبعة المعاصرة، 1974.

(The same record showing French data transcribed first)

The 312 D17 walk : a story

(Right-left data within a left-right title proper)

newly rev. and illustrated ,2 7ND

(Left-right data within a right-left edition statement)



ISBD(M) examples

Islam dan kebudayaan Melayu

= اسلام دان کيودايان ملايو .

— Kuala Lumpur : Kementerian Kebudayaan,  
Belia dan Sukan, 1977. — x, 221 ms. : 22 cm.

— (Siri kebudayaan kebangsaan ; bil. 13)

"Kertas-kertas kerja Seminar Kebudayaan 25-  
28hb Julai 1976". — Bibliografi: ms. 87-89.

— Teks Bahasa Malaysia, Inggeris dan Jawi.  
\$5.00 (kkt.)

Wahai pemuda!

= ايتها الشباب !

/ oleh Hasan al-Banna ; penterjemah Abu Bakar  
Hanzah. — Kota Bharu : Pustaka Aman Press,  
1980. — 35 ms. ; 15 cm.

Mengandungi ayat-ayat Quran.  
\$1.00 9kkt.)

سِنَاجَات دُنْيَا

= Senjata dunia

/ موسونن محمبا .

— Kuala Lumpur : Syarikat M.H.I., 1978.

— 52 موك سورة : 19 سينتي ميتر .

جودول كوليت دفن . — نام بنر فشارخ : محمد  
حسن بين اسماعيل

\$2.00 (كوليت كرتس)

IFLA Pre-Conference Seminar on Automated System for Access to  
Multilingual and Multiscript Library Materials.  
Nihon Daigaku Kaikan, August 21-22, 1986.

**MINISIS - A Multilingual Information Management System**

by

Richard Lee

Head, Future Systems

Information Sciences Division

International Development Research Centre

P.O. Box 8500,

Ottawa, Canada

K1G 3H9

Keyword(s):  
Library System;  
Information Management System;  
Non-Roman Scripts Processing.

## ABSTRACT

There are a number of computer softwares used for automating the library operations, Only few of them were designed to handle Non-Roman character sets. Because of the characteristics of Non-Roman character sets, special hardwares and specialized softwares have to be developed in order to handle the Non-Roman character sets properly. The International Development Research Centre developed a generalized information management software, MINISIS, for managing their library operations and establishing computerized bibliographic systems in developing countries. The system is capable to support both the Roman and the Non-Roman textual information. The feature of processing Non-Roman character sets is developed for the developing countries that have the needs in processing data written in their own scripts.

1(11)

## **The International Development Research Centre**

The International Development Research Centre (IDRC) is a public corporation created in 1970 by the act of the Canadian Parliament. The Centre was established to stimulate and support scientific and technical research into the problems of the developing countries. An international board of governors oversees the operations of the Centre, set the broad lines of policy and approves individual projects.

The objectives of the Centre (in the words of the Act) are "to initiate, encourage, support, and conduct research into the problems of the developing regions of the world, and into the means of applying and adapting scientific, technical, and other knowledge to the economic and social advancement of those regions" and "to help developing regions build up their own research capabilities and innovative skills needs to solve their problems." In order to carry out these objectives, it was empowered to "establish, maintain and operate information and data centres and facilities for research and other activities relevant to its objects" and "initiate and carry out research and technical development, including the establishment and operation of any pilot plan or project, to the point where the

appropriate results of such research and development can be applied."

Four major program divisions exist: Health Sciences; Agriculture, Food and Nutrition Sciences; Social Sciences; Information Sciences. The Information Sciences Division makes grants to the recipients for identifying the information needs of the developing countries; establishing information systems; processing and disseminating acquired information[1].

In 1973, IDRC decided to computerize their library operations. The ISIS (Integrated Set of Information Systems) was chosen to automate the library operations. IDRC operated the ISIS software through a service bureau. Because of the high cost for operating the ISIS software and the increasing demands for the similar software from the developing countries, IDRC decided to develop a software, MINISIS, which is operated on an inexpensive minicomputer and is functionally compatible with ISIS[2].

#### **MINISIS - A Tool for Information Management**

MINISIS is a generalized information management system which was designed to operate on the Hewlett Packard 3000 series of minicomputers. It is a member of the ISIS

family of information systems. The distinguished ability of the ISIS family is to support variable-length fields and subfields which may not exist or may repeat a variable number of times. Most of the commercial softwares do not handle efficiently with variable-length data and repeated data.

The ISIS software was created by the International Labour Office in the early 1960. The software is now maintained and distributed by Unesco(United Nations Educational, Scientific and Cultural Organization). ISIS was designed to handle bibliographic data using an IBM mainframe computers. MINISIS fills the same need except MINISIS is operated on smaller, inexpensive minicomputers. At the design stage of MINISIS, the designer of MINISIS employed a theoretical framework around the software. The relational data base theory[3] was chosen as the principle for constructing data bases. The designer also took the advantages of the advanced hardware features(stack architecture) and the sophisticated operating system to develop the software. MINISIS is equipped with a set of generalized processors which permit the users to define data bases, generate reports, retrieve data, do data entry and modification to the data bases, perform maintenance on data bases, and perform arithmetic operation on the selected data.



Although MINISIS was originally developed to permit the management of bibliographic data bases, the design of MINISIS is generalize enough for implementing a wide range of applications. It is currently operational at more than 180 installations around the world with over 100 of those in developing countries. MINISIS consists of a set of application processors. The major application processors are: ENTRY and MODIFY for entering and modifying data; QUERY for retrieving data; DATADEF for creating data bases; COMPUTE for performing arithmetic computation; PRINT for generating user-defined reports. The QUERY processor supports boolean operations, multilingual thesaurus, left and right truncation, SDI(Selected Dissemination of Information) for current awareness services. In addition, the ISOCONV processor allows the users to exchange data in ISO 2709 format, with other organizations. The system provides utilities to the data base administrator for creating and maintaining data bases.

Most of our users are able to implement their applications using the supplied application processors. However, users with special requirements can write application processors and/or specialized routines to supplement the generalized application processors provided with the system.

Among the unique feature of MINISIS is the ability to process the Non-Roman character sets. This feature enables the users to interact with the system in the preferred language and allows the users to store information in Non-Roman character sets. It is possible for users to access the same information using the same software but conducting their dialogue with MINISIS in different languages. MINISIS is being used by the organizations around the world for storing information in various combinations of English, French, Spanish, Dutch, German, Arabic, Greek, Thai, Korean and Chinese.

#### **Processing Non-Roman Character Sets**

Since the computer was invented in early 1940, computers were designed to handle data based on the Roman alphabet. Until the half of the last decade, a great deal of interest was expressed about the use of computer in processing Non-Roman scripts. Because of the advancement in the design of computer and demands in processing Non-Roman scripts, computer manufacturers have developed the input/output devices for managing the inputting/outputting of the Non-Roman character sets. Very often, the term of Non-Roman character set is referred as alternate character set by the computer manufacturers.

Based on number of characters defined in a character set, character sets can be divided into two categories, Conformable character set and Non-Conformable character set[4]. Those character sets that do not exceed 256 characters is called Conformable character set. Those character sets that exceed 256 characters is named Non-Conformable character set. Extended Roman(accented Roman alphabet), Arabic, Cyrillic are Conformable character sets. On the other hand, Chinese, Korean and Japanese are Non-Conformable character sets.

MINISIS was originally designed to handle the Roman character set. Because the users were expressed interest in processing Non-Roman character set, the feature of processing Non-Roman character set was incorporated into the system. This feature was implemented in two phases. The feature dealt with the processing of Conformable character set was added to MINISIS in the first phase. The second phase of the development of this feature dealt with the processing of Non-Conformable character set.

MINISIS supports up to 16 character sets. It is the user choices of 16 character sets. Character sets can be mixed within fields, records, and data bases. Each character set is uniquely identified with character set identification code inside the computer. A coding scheme

is employed by MINISIS for representing characters inside the computer. Most frequently, MINISIS users employ the standard coding schemes which were established by ISO (International Organization for Standardization) for Roman alphabet and some alternate character sets (ie. Cyrillic, Arabic). A code can not be used more than one time within the same coding scheme but can be shared between coding schemes. The selection of coding scheme for representing the alternate character set is totally up to the users.

Special input/output device is required for inputting/outputting Non-Roman character set. MINISIS encourages the users to use the input/output devices that can be maintained by the local computer suppliers. Beacuse of this reason, the system must be capable to work with the input/output devices that are made by different computer manufacturers. There is no standard coding scheme across the computer industry for some alternate character sets. Most frequently, computer manufacturers employ their own coding scheme for representing alternate character set. Thus problems will occur because different types of input/output devices generate different codes for the same character set and difficulties will be encountered when exchanging data and processing the character sets. An interface routine can

be written by the user for standardizing the codes of the alternate character set.

Once data are entered into the data base, various types of reports are generated from the data bases. MINISIS is equipped with a flexible report generator. The report generator handles the printing of mixed character sets as well as splitting alternate character string into lines. The printing direction of character set affects the appearance of the data on the output. Due to space required for printing a character vary between character sets, MINISIS has capability to compute number of characters fit on the output. Some alternate character set do not have clear separation between words. It is difficult to break data at the word boundary when a character string do not fit in an output line. MINISIS solves this problem by imbedding an invisible separator between words.

Sorting plays an important role in computer processing because most computer processing involves the sorting of data in some kind of sequence. Sorting the Roman alphabet is quite straight-forward but problems will be occur when sorting mixed character sets. Interfiling mixed character sets is impossible because no filing relationship exists between different character sets. Sorting characters

inside the computer is to compare the codes of characters. The smallest code is filed first and the largest code is filed last. MINISIS permits the user to establish the collating sequence for each character sets. The method for determining the character collating sequence is selected by the user. for example, our Chinese users employ either the phonetic method or stroke method for determining the Chinese character collating sequence.

MINISIS is a command-driven system. The user have to indicate the request to the application processors through commands. Then commands are parsed by MINISIS. The syntax of commands and messages used in the user dialogue are isolated from the application processors, and are stored in the external files. Users with different syntax used in commands can modify these files. Utilities provided with MINISIS are used to define the syntax of commands and alter messages to the language and the character set of the user.

#### **Availability of MINISIS**

IDRC makes MINISIS available to the users under licensing agreement. In order to maximize the resource for supporting the developing countries, IDRC has appointed commerical distributors to handle the distribution of the



MINISIS system in the commercial sector. Each commercial distributor has their own territories for marketing the MINISIS system. An initial licensing fee is required when a commercial user acquire the MINISIS system. The commercial user also pay an annual support fee for the continuing support of the system. IDRC is willing to make the system free-of-charge available to the non-profit making organizations in the developing countries. Over 100 of our 180 installations are in the developing countries. An user signed the licensing agreement directly with IDRC will receive 3-weeks training course. The training course is conducted at the user site.

11(11)

## REFERENCES

- 1 M.B. Stone, Information Systems and Services to Support the World Community (Canadian Library Journal, volume 42, number 4, August 1985), pp 203-207.
- 2 R.L. Valantin, CDS/ISIS and MINISIS: A Functional Analysis and Comparison (Ottawa, Ont.: IDRC, 1981), 88p.
- 3 F.A. Daneliuk, Information Retrieval and Library Management: and Interactive Minicomputer System (Ottawa, ONT.: IDRC, 1978) 16p.
- 4 F.A. Daneliuk and R. Lee, Computer Processing of Non-Roman Scripts, (MINISIS newsletter, Volume 1, number 2, December 1980), pp 17-24.

SIMSOVA Sylva and CLEWS John

Computers in a multi-script library environment

1986

Sylva SIMSOVA was formerly a Principal Lecturer at the School of Librarianship and Information Studies of the Polytechnic of North London where she taught computing to students of librarianship and was responsible for research projects on the reading needs of ethnic minorities. Early this year she launched an information systems consultancy Data Help (18 Muswell Avenue, London N10 2EG, tel 01-883-6351).

John CLEWS works at the British Library Document Supply Centre at Boston Spa, West Yorkshire. He actively participates in the work of the British Standards Institution's Technical Committee on character sets and information coding and the International Organisation for Standardisation working group on a 2-byte code to encompass all scripts currently used throughout the world. Last year he launched the SESAME project with the aim of removing language barriers from new technology (Sesame Computer Projects, 8 Avenue Road, Harrogate, North Yorkshire, HG2 7PG, tel 0423-888432).

The first publication of SESAME was Automating non-roman scripts by Sylva SIMSOVA and John CLEWS. Copies can be ordered by post for £ 5.

A new journal called SESAME Bulletin is due to start later this year. It will carry articles, software reviews, and news items about non-roman scripts on computers. It is planned to have a special issue on library applications.

## Computers in a multi-script library environment.

=====

The purpose of this paper is to examine the problems created by two recent trends in libraries: One is the increasing use of computers for cataloguing, information retrieval, circulation control and general administration, the other is the mix of languages libraries have to deal with. The former has been brought about by technological advancement, the latter by improvement in communications and movements of people.

## Computers and the multi-script environment.

=====

Technical inventions bear the mark of their cultural environment. The computer is no exception.

Because of the language environment in which the early computers were developed it is considered "normal" for the computer to write horizontally from left to right and to use the English version of the Roman alphabet without any diacriticals. Technical terms relating to computers are mostly English, and computer languages use English words.

The recognition of how inconvenient this language bias can be to some users dawned slowly because of the universal popularity of computers for handling mathematical tasks.

Society has a way of coping with vocabulary changes which accompany cultural borrowing. We borrowed, for instance, the term "paper" from the Egyptians, but to borrow their hieroglyphic script as well would have been no advantage to English.

As computer technology stands, computer users cannot make this distinction. If their method of writing is from right to left or vertical, and the shapes of their alphabet differ from the Roman one, they find themselves in a dilemma: Should they forgo the advantages of computer technology or should they forgo their own language.

The spread of a technical invention can be likened to an epidemic of cultural borrowing. Because of their convenience technical inventions cross borders rapidly. Their recipients do not have time to ponder the implications of the cultural discrepancies between the two environments.

The convenience of having a tool for large computations and for running systems effectively cancels out the inconvenience

of learning a foreign language for programming and a foreign script for text handling.

It takes some time before the users, having become familiar with the new tool, realise that this inconvenience is unnecessary and begin to voice demands that the computer should fit in better into their own linguistic environment. There is, for instance, hardly any mention in Indian professional literature that the use of computers in Indian libraries is limited by the Roman script.

Computer experts are more vocal than end users in this matter. An increase in the number of computers throughout the world means an increase in the number of computer experts whose first language is not English. In recent years, therefore, the number of articles asking for multiscript computers has grown considerably.

Once the need has been pointed out and is noticed by commercial entrepreneurs the development is rapid. We have just now entered a period when new multi-script developments abound.

#### Growth of interest in non-roman scripts

=====

An analysis of the entries in a recently published bibliography on "Automating non-roman scripts"[1] suggests three successive stages in the development of interest in the subject. The items are mostly research reports or general articles in English language computing and librarianship journals.

The first stage covers the period of 1959 to 1974. Apart from two early references to Chinese in 1959 and 1963 the interest did not really start until 1970 with an average of 1 reference per year. Half of the references dealt with Chinese and half with Devanagari.

The second stage covers the period of 1975 to 1979 with an average of 5 references per year. About half of them dealt with the Far Eastern scripts (8 Chinese, 6 Japanese), a quarter with South and South East Asian scripts (4 Devanagari, 1 Thai, 1 Tamil), a fifth with the versions of the Arabic script (3 Arabic, 2 Persian) and the remainder with Hebrew and Cyrillic (1 each).

The third stage covers the period of 1980 to 1985 with an average of 32 references per year. About three quarters of them dealt with the Far Eastern scripts (11 Far Eastern in

general, 88 Chinese, 46 Japanese, 3 Korean), one seventh with South and South East Asian scripts (5 Indic in general, 4 Devanagari, 6 Tibetan, 3 Thai, 2 Gurumukhi, 2 Khmer, 1 Tamil, 1 Telugu), one fourteenth with the versions of the Arabic script (12 Arabic, 2 Urdu, 1 Persian), and the remainder with Hebrew (2), Cyrillic (2), Greek (1).

The bibliography on which this analysis is based includes only material in English. Research literature on non-roman scripts is now increasingly appearing in other languages, in particular Japanese. A quick count of Japanese material accessible in libraries in London suggests the following pattern of growth of Japanese interest:

The first stage covers the period of 1973 to 1979. Although there are two early references in 1973 and 1977 the interest seems to have started around 1978 with an average of 3 references per year for 1978 and 1979.

The second stage covers the period of 1980 to 1985 with an average of 7 references per year.

The growth of interest could also be plotted out by analysing the promotion of commercially available equipment. Publicity material collected over the last year from computer exhibitions and conferences offers to the user a confusing picture of abundance on one hand and poverty on the other. Casual enquiries to general computer dealers are frequently met with a blank stare. The various versions of the Roman script with their diacritical marks are now generally available, but a request for non-roman scripts can take some dealers by surprise. On the other hand there are advertisements which suggest that more than one system exist for many scripts and there is evidence that developments are taking place which remain unpublicised either in research literature or the commercial channels.

Research activity in this subject is badly documented because much of it is carried out by enthusiasts without proper funding. To illustrate from one of the less used scripts:

In 1980 there was only one system for computer generation of the Tibetan script.[2] It was purely a research case study and was not commercially available for practical application. In 1983 it was followed by an alternative system designed by one of the authors of this paper, also as a research case study only.[3]

Tibetan is a little known language and the existence of two computer systems may come as a surprise. Today, however, there are at least ten people in various parts of the world who are reported to be working on the problem.



One American package is commercially available, two American research projects have been reported in literature, one in an engineering journal and another in a periodical devoted to Tibetan Buddhism. One of them can be obtained on application presumably free or for minimal cost. In France there is a system designed for practical application in a fairly advanced stage, and in Japan there is another one still being developed. In Britain there are at least three systems in an early stage of development. [4] A recent article in the Tibetan Review called for standardisation of a Tibetan character code to overcome the problem of multiplicity of systems. [5]

#### Libraries in a multi-script environment

Many libraries now have to handle books and periodicals in more than one script, particularly major research libraries with large oriental collections, and public libraries serving Asian ethnic minorities, as well as libraries in countries which principally use one or more non-roman scripts.

Early libraries in the Western world contained material in three scripts at the most: Roman, Greek, Hebrew.

With the growth of Oriental scholarship and increasing contacts among the countries of Europe, multilingual collections were created containing material in Oriental languages, in versions of the Roman script, and in Cyrillic.

Some libraries catalogued material in each script separately so that entries for each script could be arranged according to the rules of the script. Others transliterated all scripts into Roman so that all entries could be in one sequence. A variant of the latter method was to include the original script side by side with the romanisation either for the full entry or its parts, e.g. the title.

Nearly all non-roman scripts used a much richer repertoire of symbols than roman script, but library transliteration schemes forced them into a procrustean bed of 26 letters.

One of the problems was that scripts could be romanised in more than one way. In spite of attempts at codification of the transliteration rules and standardisation of those that had been codified, the same non-roman titles were recorded in diverse forms in the catalogues of different libraries.

As long as the number of libraries with multi-script stock was small and their catalogues were locally produced by hand

this did not pose any great difficulties except to the few scholars who used more than one library but who soon got used to the local rules of each of them.

The problem has, however, been highlighted in the last decades by three developments:

The first is the trend towards library cooperation and universal bibliographic control. Libraries are no longer seen as completely separate entities, but as part of a system.

The second is the spread of library automation which makes it technically inconvenient or impossible to keep local catalogues in the original script since the computer is geared, on the whole, to the English version of the Roman script. Some libraries accept transliterated automated records while continuing to keep manual ones in the original script, but most resign themselves to the loss of valuable information carried by the original script. It is to be hoped that this is only a passing phase due to the shortcomings of the technology available at the time.

The third is the change in the type of libraries with multi-script material. Oriental books in the West are no longer only supplied by scholarly libraries to meet the needs of academics, nor are those in the Greek script just read by the students of classics. Due to the movement of people throughout the world the linguistic composition in most large cities has become complex. There are potential readers of a wide range of scripts among the traditional users of the public library. Multi-lingual catalogues in public libraries are a new development. In Britain there have always been a few catalogue entries for a sprinkling of German or French novels in public libraries, but apart from these the stock has been predominantly monolingual. There is therefore no tradition of cataloguing material in non-roman scripts and there is usually little contact with academic libraries who have longer experience of this problem. To the overworked public librarian transliteration into the Roman script offers an easy solution overlooking the fact that, for instance, to a Cantonese a Romanised Chinese title is as good as lost.

#### Non-roman script readers in London

=====

As librarians we lack information about the problems of the non-roman script reader. We need more research into the needs of such readers in order to use the available resources wisely.

The needs of non-roman script readers are often met by self-

help without librarians knowing about it. A recent survey of reading-related activities by the linguistic minorities in the London area shows an "alternative" resource of material in a variety of languages existing side by side with the academic and public libraries.[6]

Last December a sample of self-help organisations listed in the report of this survey was telephoned to find out whether they were using computers. A successful contact was made with 20 self-help organisations of whom 9 were South Asian, 7 Moslem, 2 European (Greek, Ukrainian), 1 East Asian, 1 Chinese and 1 Japanese.

They were asked what methods they were using to produce texts in their own language. The most common method was the use of a typewriter with a non-roman script (9). About a third used various methods of word processing or computerised typesetting (7), a few used calligraphy (4) or typesetting abroad (3) or typesetting locally (2).

The South Asian self-help institutions used the typewriter (5) or computerised methods of text production (4) about equally. A few also used calligraphy (2) or had their text typeset in India (2).

The Moslem ones used mostly a typewriter (3), one used computerised typesetting and one conventional typesetting.

Of the East Asian ones two used calligraphy, one a typewriter and one had text typeset in Tokyo.

The Greek one used typesetting and the Ukrainian one computerised typesetting.

None of the self-help institutions were using any form of library automation, although some of them had libraries. Those who did use any form of automation were mostly involved in publishing or production of leaflets and other ephemera.

To give a few examples : the Gujarati Samachar (a newspaper) and the Ukrainian Publications were both printed by computerised linotype. There was at least one firm which was offering computerised typesetting in 73 languages.[7]

It seems that at the time of the survey, i.e. end of 1985, the linguistic minorities and their host society had access to relatively cheap computerised typesetting and were beginning to use it. The use of computers for word processing was a different matter. Typewriters and calligraphy were still the most common methods for text production.

The reason was twofold: The equipment for word processing in

non-roman scripts was not easily available and linguistic minorities, in common with the host society, were lacking in computer literacy.

To give the minorities a headstart in computer literacy a project has now been started by the London Voluntary Service Council. It is called "Computer Access for London Communities". One of the units is going to be located in Ealing which has a large number of South Asians. [8]

There are signs that this is the beginning of a new trend. To quote two other examples: The Information Technology Centre in Bradford, another area of South Asian settlement, is developing software in non-roman script. [9] In London, a tutorial package has been developed for learning written Chinese. [10]

These two examples indicate the directions in which future developments may go: One is the provision of technology for non-roman scripts on computers, the other a social acceptance of the computer as a tool among the ethnic minorities.

#### Multiscript computers in libraries

=====

On the whole, libraries have up to now been concentrating on other aspects of automation and relying on transliteration for their cataloguing.

A decade ago Wellisch [11] pointed out that transliteration has two aims which tend to be incompatible. These are script reversibility, so that the state of the original script can be shown and pronounceability, showing the sound of the original language. Although transliteration into a uniform script enables document descriptions to be filed in a single sequence, much of the information content of the original script is lost in transliteration.

Today, a computerised catalogue can make use of the filing capabilities of computers to arrange the whole sequence in one predominant script, as well as offering additional alphabetic sequences in other scripts.

The International Organisation for Standardization committee responsible for transliteration now aims to produce completely reversible transliteration schemes, so that if they are adopted in computerised library catalogues it should be easy to arrange computerised display of the original script at some time in the future.

In the United Kingdom, Cambridge University Library has

devised a transliteration scheme for Hebrew which allows input and display using the characters of the USMARC character sets, but which enables separate printouts to be produced in Hebrew and Arabic script, via a separate printer. However, for Arabic materials, input is a time-consuming two-stage process of Arabic to Roman script conversion for the cataloguer, and Roman to Arabic script for the computer.

The most useful approach for cataloguers involves a one-stage input in the original script. This has recently been achieved in Israel for Hebrew materials [12], where a comparatively limited character set of 22 Hebrew consonants is used.

Much harder problems of vernacular script input have been achieved in Japan, where thousands of Chinese characters (Kanji) are in daily use. Japan Information Centre for Science and Technology (JICST) provides the online data base system JOIS and several abstracting journals, indexed by Japanese phonetic symbols (Kana). The National Diet Library of Japan also produces JAPAN-MARC exchange tapes using Kana and Kanji, and JAPAN-MARC records are also available through the Canadian "bibliographic utility" UTLAS, based in Toronto.

Outside Japan it is the large North American bibliographic utilities where most developments have taken place. The RLG/CJK terminal, developed from a Transtech terminal by the Research Libraries group has been the most widely documented system. This uses a character-component input method which appears to the user as if he is "composing" each character from its constituent parts, many of which are the same as the traditional Chinese radicals used in China, Japan and Korea.

While this method allows in theory the composition of any character in Chinese, Japanese or Korean, composing the characters is slow. It may only take an average of 3.7 keystrokes to "compose" a character, but it may take several seconds for each component part to be found on the keyboard. In addition, processing takes place at a remote database, with no local processing possible. Despite these shortcomings, the RLG achievement has enabled thousands of CJK records to be produced for exchange by RLG member libraries.

The 2-byte REACC code used by the RLG/CJK terminal is also used by OCLC for its own CJK system, due out in 1986, which has a character by character phonetic input method, using a QWERTY keyboard, developed by Asiagraphics. OCLC also offers some local processing facilities, at least for its Roman script records.

Outside North America, China and Korea have been independently developing their own MARC record systems, each



using their own 2-byte code. China plans to distribute exchange tapes during 1986. In Taiwan, exchange tapes have been produced for some time, using the 3-byte code CCCII, which is a subset of the REACC 3-byte code. The National Central Library in Taiwan will also be contributing CJK records via OCLC.

#### Gaps in multiscrypt computing in libraries

=====

Despite these advances, there are still several areas which need further attention if computers in libraries are to take full advantage of developments in computers for other applications.

Firstly, although phonetic input systems are available for library computers, none of these offer a touch typing facility to speed input, unless some of the Japanese systems use such terminals.

Secondly, few local processing systems use a character set other than the extended roman USMARC character set, or the extended roman ISO 5426 set applied in UNIMARC and some other national character sets for libraries. The DOBIS/LIBIS system running on IBM equipment does offer an Arabic character set for its major Arabian clients; also some libraries in Israel use a mixed Latin/Hebrew character set, with Hebrew characters in place of lower case ASCII characters; the cooperative SEAPRINT programme of CONSAL intends to include Thai characters as well as Roman script used by other CONSAL members. RLG plans to add Hebrew and Cyrillic to the CJK capabilities. However, apart from these isolated instances, a large number of Asian scripts - particularly in South and South East Asia - are not available on any library computer. In the commercial world, at present only one company - Xerox - offers most world scripts on one set of equipment.[13] In commerce, there are few multiscrypt applications - most companies frequently subcontract documentation involving specific non-roman scripts such as Japanese or Arabic to specialist translation companies. Library catalogues are the major multiscrypt environment, but there is no multiscrypt computer or terminal to aid library automation.

Thirdly, there is a gap in the area of standards for character sets. For instance, China, Japan and Korea have completely different 2-byte national character sets, and in Taiwan an unrelated 3-byte code is used. However, there is at least no inconsistency between library and commercial applications as exists in the West: In the ISO there is no liaison between the international documentation and the computer committees, though each of them liaises with its



national counterparts. Consequently a number of incompatible character sets are in use.

Unfortunately, the most widely used character set in libraries - the USMARC character set - is not a standard at all, although it appeared in an annex to a standard as early as 1969, and a related standard appeared in 1985. Neither was it registered under the provisions of ISO 2375, so it could not be used as a character set within UNIMARC.

Fortunately there have been signs of willingness to tackle some of these standardisation problems. In particular British members of the computer fraternity and the library community have recently shown a much greater interest in liaison both in BSI and within working groups of ISO, and it is hoped that this will rapidly lower the number of incompatible character sets.

Furthermore, REACC, despite being yet another character set for Far Eastern scripts, has great potential for compatibility with all of them. It is totally compatible with the Taiwanese CCCII, and potentially compatible with the Chinese GB-2312 and the Japanese JISC-6226.

#### The future =====

Librarians have, on the whole, been slow in recognising the problems caused by the dominance of the roman script in their records to readers of non-roman scripts. There are signs that this is now changing as efforts are being made to reconcile the pressure of users' needs on the one hand and of commercial products on the other. It is likely that in the near future, multi-script records will be a commonplace.

To assist this development, further work needs to be done in three directions simultaneously: technological enhancements of hardware and software, keeping librarians informed about this field, and spreading computer literacy among non-roman script readers.

The theme of this IFLA pre-conference and the fact that it brings together specialists in information technology and in the needs of linguistic minorities augurs well for the future.

## References

=====

- [1] SIMSOVA Sylva and CLEWS John comp. Automating non-roman scripts. Sesame, 1986.
- [2] NANCARROW Peter. A system for processing Tibetan texts in their original orthography. ALLC Journal, 1(1)1980, 19-24.
- [3] SIMSOVA Sylva. Multilingual information and the computer. Information Development 1(2)Apr 1985, 103-8.
- [4] The Tibetan Text Processing System can be ordered from M.S.T.P. 216a West Second St, Howell, NJ 07731.  
ZHANG Liensheng. A design for the computerized processing of Tibetan scripts. Journal of the Institute of Electronic and Telecommunication Engineers 3096)Nov 1984, 257-60.  
ROCKWELL John. Carving Tibetan in silicon: translation, word processing and computers. The Vajradhattu Sun, Oct/Nov 1985, .  
Monsieur Christian Bodson, France.  
Mr Ito, Japan.  
Michael Mann, SOAS, London.  
Edward Henning, London.  
Peter Lofting, London.
- [5] WOOLSEY Jim. Tibetan in computer. Toibetan Review 21(2) Feb 1986, 19.
- [6] ELLIOTT Pirkko. Public libraries and self-help ethnic minority organisations. The Polytechnic of North London, 1984. (SOLIS Research Report No 12)
- [7] WARD Gareth. Mixed reraction to Mapex. Printing World 214(16)23 Oct 1985, 27. (Merton Graphics)
- [8] C.A.L.C - Computer access for local communities. Computanews (20)1985, 1.  
CALC choses host groups. Computanews (21/2)1985, 3.
- [9] Asian script word processing in Bradford. The Incorporated Linguist 24(7)Spr 1985, .
- [10] Paul Thompson, SOAS, London.
- [11] WELLISCH Hans Hanan. The conversion of scripts. Wiley, 1977.
- [12] ADLER Elhanan. Judaica automation in Israel - an overview. Judaica Librarianship 1(1)Fall 1983, 9-11.
- [13] BECKER Joseph D. Multilingual word processing. Scientific American 151(1)Jul 1984, 96-107.